VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ
BRNO UNIVERSITY OF TECHNOLOGY

FAKULTA INFORMAČNÍCH TECHNOLOGIÍ
ÚSTAV POČÍTAČOVÝCH SYSTÉMŮ

FACULTY OF INFORMATION TECHNOLOGY
DEPARTMENT OF COMPUTER SYSTEMS

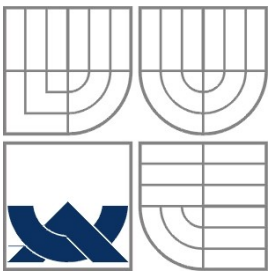
DETEKCE GENŮ V DNA SEKVENCÍCH

BAKALÁŘSKÁ PRÁCE
BACHELOR'S THESIS

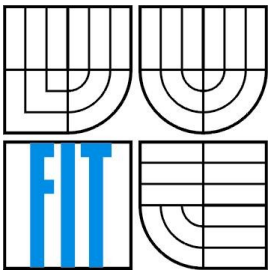
AUTOR PRÁCE
AUTHOR

TOMÁŠ BAHUREK

BRNO 2011



VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ
BRNO UNIVERSITY OF TECHNOLOGY



FAKULTA INFORMAČNÍCH TECHNOLOGIÍ
ÚSTAV POČÍTAČOVÝCH SYSTÉMŮ

FACULTY OF INFORMATION TECHNOLOGY
DEPARTMENT OF COMPUTER SYSTEMS

DETEKCE GENŮ V DNA SEKVENCÍCH

GENE DETECTION IN DNA SEQUENCES

BAKALÁŘSKÁ PRÁCE
BACHELOR'S THESIS

AUTOR PRÁCE
AUTHOR

TOMÁŠ BAHUREK

VEDOUCÍ PRÁCE
SUPERVISOR

Ing. TOMÁŠ MARTÍNEK, Ph.D.

BRNO 2011

Abstrakt

Tato práce se věnuje problematice detekce genů v DNA sekvencích. Vysvětluje spůsob uložení informace v DNA sekvencích, poskytuje přehled některých metod pro detekci genů, konkrétně popisuje vybrané metody a výsledky testů těchto metod na reálnych datech (geném bakterie E. coli K12). Obsahuje porovnání efektivity metod při různých parametrech a efektivity metod mezi sebou.

Abstract

This work is about gene detection in DNA sequencies. It explains how the information in DNA is stored, shows preview of various gene prediction methods, describes picked methods and results of testing those methods on real data (genome of bacteria E. Coli K12), compares efficiency of methods with various paramaters and efficiency between methods.

Klíčová slova

molekulární biologie, DNA, detekce genů, génová predikce, prokaryoty

Keywords

molecular biology, DNA, gene detection, gene prediction, prokaryotes

Citace

Tomáš Bahurek: Detekce genů v DNA sekvencích, bakalářská práce, Brno, FIT VUT v Brně, 2011

Detekce genů v DNA sekvencích

Prohlášení

Prohlašuji, že jsem tuto bakalářskou práci vypracoval samostatně pod vedením Ing. Tomáš Martínek. Uvedl jsem všechny literární prameny a publikace, ze kterých jsem čerpal.

……
Tomáš Bahurek
14.5.2011

Poděkování

Chcel by som poďakovať vedúcemu tejto práce, Ing. Tomášovi Martínkovi za jeho užitočné rady a za to, že ma vždy naviedol správnym smerom.

Obsah

1 Úvod

Gény sú základom dedičnosti v organizmoch a podkladom na vytvorenie proteínov. V génoch sú informácie ako vytvoriť a udržovať bunky organizmu a ako preniesť genetické znaky na potomstvo. Organizmy majú mnoho génov, ktoré ovplyvňujú rôzne biologické znaky. Niektoré znaky sú viditeľné (farba očí, počet končatín...) a niektoré nie (typ krvi, náchylnosť k určitému typu chorôb...). Gény sú tie časti DNA, ktoré nesú genetickú informáciu. DNA však pozostáva nielen z génov, ale aj z častí, ktoré majú buď štruktúrne účely alebo sa podieľajú na regulácii používania genetickej informácie. Kedysi bola detekcia génov založená na experimentoch s bunkami. Experimentami boli zmapované gény niektorých organizmov a boli nájdené určité opakujúce sa vzory, podľa ktorých je možné gén detekovať. Vďaka týmto informáciám a neustálemu zdokonaľovaniu výpočtovej techniky sa vyhľadávanie génov v dnešnej dobe rieši prevažne pomocou algoritmov na detekciu génov. Detekcia génov je nevyhnutná k predikcii štruktúry proteínu a pochopeniu jeho funkcie. Rozpoznanie génov je teda nevyhnutné v pochopení fungovania akéhokoľvek organizmu.

Cieľom tejto bakalárskej práce je oboznámenie sa so základnými princípmi molekulárnej biológie, spôsobmi uloženia informácie v DNA sekvenciách a algoritmami pre detekciu génov v DNA sekvenciách a návrh vhodnej metódy pre detekciu génov.

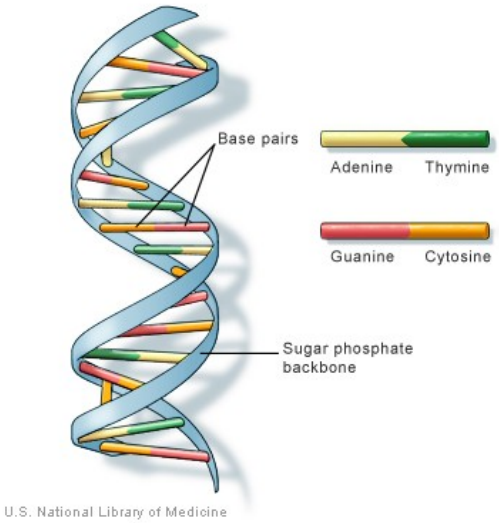
Práca je rozdelená do 6 kapitol. Hneď po úvode nasleduje kapitola **Molekulárna biológia**, kde sú vysvetlené základné pojmy, spôsom uloženia informácií v DNA sekvenciách a z neho vyplývajúce problémy v detekcii génov. V kapitole **Metódy génovej predikcie** sú teoreticky popísané niektoré metódy, s tým, že detailnejší popis majú väčšinou metódy, ktoré boli aj implementované v rámci tejto práce. Kapitola **Presnosť Génovej predikcie** vysvetľuje spôsoby hodnotenia jednotlivých metód. V kapitole **Experimeny** sú rozoberané výsledky jednotlivých metód s rôznymi parametrami. Na konci tejto kapitoly sú jednotlivé metódy porovnané medzi sebou. Kapitola **Záver** obsahuje zhodnotenie práce a možnosti jej pokračovania

2 Molekulárna biológia

Molekulárna biológia je vedný obor zaoberajúci sa živými organizmami na úrovni molekúl. Jeho pole pôsobnosti sa prekrýva s inými oblasťami biológie a chémie, najmä s genetikou a biochémiou. Molekulárna biológia sa zaoberá predovšetkým porozumením vzťahov medzi rôznymi bunkovými systémami, zahŕňajúc napríklad vzájomné vzťahy medzi DNA, RNA, proteosyntézou (tvorbou bielkovín), biosyntézou enzýmov, porozumením ako sú tieto interakcie regulované a skúma aj zákonitosti genetickej regulácie a genetický kód. Vysvetlenie pojmu DNA, centrálnej dogmy, transkripcie a translácie bolo spracované zo zdrojov [4] a [10].

2.1 DNA

Deoxyribonukleová kyselina (DNA) je dedičný materiál vo väčšine živých organizmov. Takmer každá bunka v tele človeka má rovnakú DNA. V eukaryotických bunkách je uložená v jadre, ale malé množstvo DNA môžeme tiež nájsť v mitochondriách a chloroplastoch. DNA je nositeľom genetickej informácie, ktorá je v DNA je uložená ako kód pozostávajúci zo 4 chemických báz: *adenín (A)*, *guanín (G)*, *cytozín (C)* a *tymín (T)*. Ľudská DNA pozostáva z približne 3 miliárd báz a viac ako 99 percent týchto báz je rovnakých u všetkých ľudí. Sekvencia týchto báz obsahuje informácie pre budovanie a spravovanie organizmu.



DNA bázy sa spájajú do dvojíc vodíkovými väzbami. A sa spája s T, C sa spája s G, čím vznikajú takzvané bázové páry. Každá báza je tiež napojená na molekulu cukru a molekulu fosfátu. Báza, cukor a fosfát tvoria dokopy nukleotid. *Nukleotidy* sú usporiadané v dvoch komplementárnych (nie však identických) vláknach, ktoré vytvárajú dvojitú špirálu.

Dôležitou vlastnosťou DNA je, že sa dokáže replikovať. Každé vlákno DNA v dvojitej špirále môže slúžiť ako vzor pre duplikovanie sekvencie báz. Toto je dôležité pri delení buniek, pretože každá nová bunka musí obsahovať kópiu DNA z pôvodnej bunky.

2.2 Centrálna dogma molekulárnej bológie

Gén je v klasickej genetike základnou jednotkou dedičnosti. Gény sú prevažne uložené v DNA v chromozómoch. Gény sú časti DNA, ktoré slúžia ako podklad na vytváranie *proteínov*. Gény u človeka môžu mať dĺžku od stoviek nukleotidových báz až po viac ako 2 milióny báz. Výskumy

odhalili funkciu niektorých génov (niektoré majú napríklad súvislosť s náchylnosťou k určitej chorobe), u mnohých génov je však funkcia zatiaľ neznáma.

Proces vytvorenia proteínu z génu pozostáva z *transkripcie* a *translácie*. Tejto dvojici procesov sa hovorí *génová expresia*.

2.3 Transkripcia a translácia

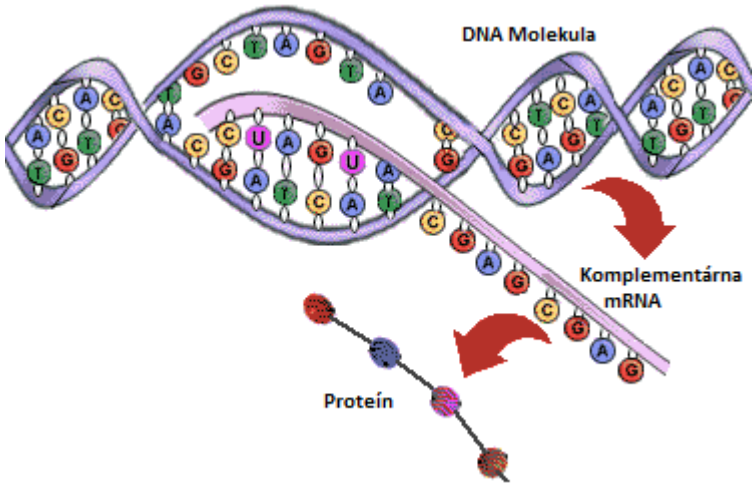
RNA aj DNA sú obe z reťazca nukleotidových báz, ale majú odlišné chemické vlastnosti. Typ RNA, ktorý obsahuje informáciu na vytvorenie proteínu sa nazýva mRNA (mediátorová RNA, v angličtine messenger RNA), pretože nesie informáciu (správu) z DNA v jadre do cytoplazmy.

Transkripcia je prepis reťazca RNA podla vzorového reťazca DNA, pričom jednotlivé nukleotidy sú pripojované na základe komplementarity (viz predošlá podkapitola). Kľúčovým enzýmom tejto syntézy je RNA-polymeráza. Prebieha v troch stupňoch:

Iniciácia, kedy sa RNA-polymeráza viaže na špecifickú sekvenciu DNA a presunuje k miestu, kde začína vlastná syntéza.

Elongácia, kedy sa RNA-polymeráza posúva ďalej pozdĺž reťazca DNA, uvoľňuje kódujúci reťazec a podľa vzorového reťazca postupne syntetizuje novú RNA tým, že na volnú 3´-OH skupinu ribózy pripojuje komplementárne nukleotidy. Vznikajúca RNA sa postupne uvoľňuje od komplexu s DNA a dvojitá špirála DNA sa samovoľne obnovuje.

Terminácia (ukončenie syntézy a uvoľnenie RNA) je signalizovaná zvláštnymi sekvenciami v štruktúre DNA, ktoré sú rozpoznávané bielkovinami, tzv. terminačnými faktormi.

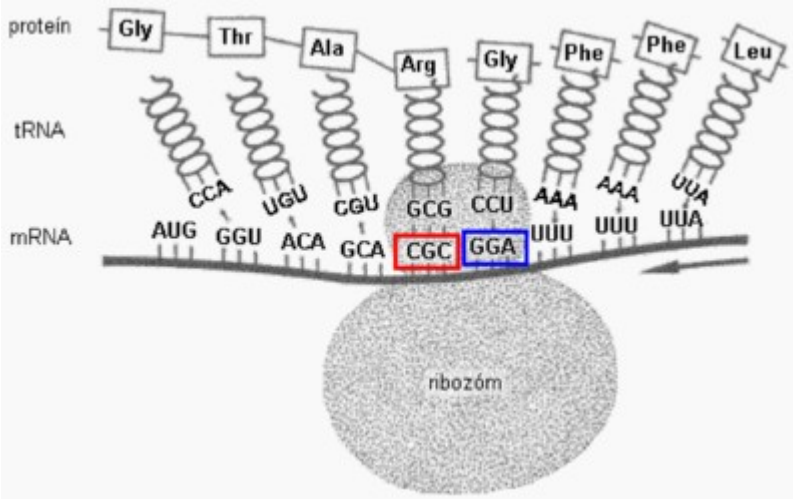


Transkripcia (prevzaté z www.scientificpsychic.com)

Translácia sa odohráva v cytoplazme. V tomto procese mRNA interaguje s ribozómami, ktoré "čítajú" sekvenciu báz v mRNA. Každá trojica báz (napr. AUG,CGG,CAG...) zvyčajne kóduje jednu konkrétnu aminokyselinu (aminokyseliny sú stavebnými jednotkami proteínu). Takáto trojica báz sa označuje pojmom *kodón* (zoznam všetkých kodónov je možné nájsť v tabuľke 1). Druh RNA, ktorý sa nazýva tRNA (transferová RNA) zostaví proteín z jednotlivých aminokyselín. Tvorba proteínu prebieha, kým ribozóm nenarazí na *stop kodón* (sekvencia troch báz, ktoré už nekódujú aminokyselinu).

	kodón	aminokyselina	kodón	aminokyselina	kodón	aminokyselina	kodón	aminokyselina
	TTT	Phe(F)	TCT	Ser(S)	TAT	Tyr(Y)	TGT	Cys(C)
	TTC	Phe(F)	TCC	Ser(S)	TAC	Tyr(Y)	TGC	Cys(C)
	TTA	Leu(L)	TCA	Ser(S)	TAA	STOP	TGA	STOP
T	TTG	Leu(L)	TCG	Ser(S)	TAG	STOP	TGG	Trp(W)
	CTT	Leu(L)	CCT	Pro(P)	CAT	His(H)	CGT	Arg(R)
	CTC	Leu(L)	CCC	Pro(P)	CAC	His(H)	CGC	Arg(R)
	CTA	Leu(L)	CCA	Pro(P)	CAA	Gln(Q)	CGA	Arg(R)
C	CTG	Leu(L)	CCG	Pro(P)	CAG	Gln(Q)	CGG	Arg(R)
	ATT	Ile(I)	ACT	Thr(T)	AAT	Asn(N)	AGT	Ser(S)
	ATC	Ile(I)	ACC	Thr(T)	AAC	Asn(N)	AGC	Ser(S)
	ATA	Ile(I)	ACA	Thr(T)	AAA	Lys(K)	AGA	Arg(R)
A	ATG	Met(M)	ACG	Thr(T)	AAG	Lys(K)	AGG	Arg(R)
	GTT	Val(V)	GCT	Ala(A)	GAT	Asp(D)	GGT	Gly(G)
	GTC	Val(V)	GCC	Ala(A)	GAC	Asp(D)	GGC	Gly(G)
	GTA	Val(V)	GCA	Ala(A)	GAA	Glu(E)	GGA	Gly(G)
G	GTG	Val(V)	GCG	Ala(A)	GAG	Glu(E)	GGG	Gly(G)
	T		C		A		G	

Tabuľka 1: Všetky kodóny a k nim prislúchajúce aminokyseliny



Translácia (prevzaté z sk.wikipedia.org)

2.4 Signály

Hlavnú rolu pri transkripcii zohráva *RNA Polymeráza*, ktorá nasadá na oblasť *promotéru* – špecifická oblasť cca 30 báz pred začiatkom génu. K transkripcii dochádza iba za predpokladu, že sú v oblasti promotéru prítomné aj príslušné *regulátory* – proteíny, ktoré rozhodujú či k transkripcii dôjde alebo nie (negatívny regulátor – zabraňuje transkripcii, pozitívny regulátor – bez neho k transkripcii nedôjde).

2.4.1 Štartovacie signály

Štruktúra promotéru:

- Transkripcia začína na ofsete 0
- Pribnow Box (TATA box) začína na ofsete -10
- Gilbertov Box začína na ofsete -30

Pribnow Box:

sekvencia	***T***	***A***	***T***	***A***	***A***	***T***
Pravdepodobnosť (%)	79,00%	95,00%	44,00%	59,00%	51,00%	96,00%

Gilbertov Box:

sekvencia	***T***	***T***	***G***	***A***	***C***	***A***
Pravdepodobnosť (%)	82,00%	84,00%	79,00%	64,00%	53,00%	45,00%

U jednotlivých sekvencií je uvedená pravdepodobnosť výrazu. Niekedy sa uvádza aj *logo* alebo *konsenzus sekvencia* (podobná regulárnemu výrazu).

Charakteristickým znakom prokaryotických génov je, že obsahujú takzvané *Shine-Delgarno sekvencie*. Tieto sekvencie sú umiestnené medzi začiatkom transkripcie a štart kodónom a pomáhajú ribozómom začať preklad. Existuje viac typov týchto sekvencií, ale väčšina obsahuje znaky: AGGAGGU [6].

2.4.2 Gilbertov a Pribnow Box

Teória k tejto podkapitole bola spracovaná zo zdroja [12]. V roku 1975 David Pribnow objavil sekvenciu piatich RNA polymeráz. Identifikoval sekvenciu sústreďujúcu sa 10 bázových párov pred začiatkom transkripcie. Sekvencia sa pôvodne nazývala *Pribnow box*. Dnes je častejšie označovaná ze *-10 región*.

35 bázových párov pred začiatkom transkripcie sa nachádza sekvencia TTGACA

10 bázových párov pred začiatkom transkripcie sa nachádza sekvencia TATAAT

Vzdialenosť medzi týmito sekvenciami zvykne byť 17 ± 1 bázových párov.

TTGACA ---- 17±1 ----TATAAT

Táto sekvencia bola časom potvrdená. Zistilo sa však, že v nej nemusia byť obsiahnuté všetky bázy.

Percento zachovania každého bázového páru v kompilácii od Harleyho a Reynoldsa založenej na analýze 263 promotérov u baktériofágov, plazmidov a baktérií:

$$T_{78}T_{82}G_{68}A_{58}C_{52}A_{54}\text{ -- }16_{21}17_{52}18_{19}\text{ -- }T_{82}A_{89}T_{52}A_{59}A_{49}T_{89}$$

Táto kompilácia a mnoho ďalších boli kritizované pre určité skreslenie v týchto smeroch:

- je ľahšie skúmať silné promotéry než slabé
- promotéry baktériofága sú väčšinou silné a preto nie sú až tak reprezentatívne v prípade E. Coli baktérie

Z tohto dôvodu Shlomit Lisser a Hanah Margalit skúmali výhradne promotéry v E. Coli. Ich výsledky ukázali zmeny v percentách výskytu niektorých báz.

$$T_{69}T_{79}G_{61}A_{56}C_{54}A_{54}\text{ -- }16_{17}17_{43}18_{17}\text{ -- }T_{77}A_{76}T_{60}A_{61}A_{56}T_{82}$$

Význam variability v promotérovej sekvencii

Prečo je možné tak veľké množstvo variability v promotéry baktérie E. Coli? Zatiaľčo RNA polymeráza u E. Coli je určená k transkripcii mRNA, nie všetky mRNA molekuly musia byť vytvorené v rovnakom množstve. Najjednoduchším spôsobom ako kontrolovať úroveň mRNA syntézy je rôzniť promotérové sekvencie tak, že RNA polymeráza rozpozná niektoré veľmi dobre a niektoré menej.

Preto existujú silné a slabé promotéry:

Silný promotér

Promotér recA je silný promotér, pretože sa líši od vzorového E. Coli promotéra len jedným nukleotidom a má o jednu bázu kratšiu medzeru.

TTGATA -- 16 -- TATAAT

TTGACA -- 17 -- TATAAT

Slabý promotér

AraBAD promotér je slabý promotér. Tento promotér sa líši od vzorového promotéra 5 nukleotidmi a 1 bázou v dĺžke medzery.

CTGACG -- 18 -- TACTGT

TTGACA -- 17 – TATAAT

2.4.3 Väzobné miesta ribozómov

Väzobné miesta ribozómov (anglicky Ribosomial Binding sites, skratka: RBS) je oblasť, kde sa ribozóm viaže na mRNA, aby začal transláciu mRNA do proteínu. V dostupnej literatúre je možné nájsť veľké množstvo definícii pre RBS, ale niektoré charakteristiky sú spoločné:

- RBS sekvencie sú bohaté na Adenín (A) a Guanín (G)
- sú lokalizované 3 až 14 bázových párov pred začiatkom génu
- ich dĺžka sa môže meniť od 3 do 9 bázových párov
- bežne sa uvádza konsenzus sekvencia AGGAG

[7]

2.4.4 Ukončovacie signály

Drvivá väčšina génov (viac ako 90%) obsahuje aj ukončovacie sekvencie pre transkripciu – *terminátory*. Pre terminátory je charakteristické, že obsahujú *palindromatickú štruktúru* (je z oboch strán rovnaká) dĺžky *7-20 nukleotidov* nasledovanú cca *6 uracilmi (U)*. Palindromatická štruktúra u RNA vytvorí pevnejšie väzby medzi GC a AU a vytvorí takzvanú sekundárnu štruktúru RNA. Experimentálne bolo zistené, že ak RNA polymeráza narazí na túto štruktúru, tak sa pozastaví na cca 1 minútu. To je obrovské spomalenie oproti bežnej rýchlosti 100 nukleotidov za sekundu a spôsobí, že väzby uracilu a potenciálneho adenínu už nie sú tak silné a transkripcia sa ukončí [6].

2.5 Intróny

Intrón je časť génu, ktorá nekóduje aminokyselinu. V bunkách eukaryotov, je väčšina génových sekvencií prerušených jedným alebo viacerými intrónmi. Časti génovej sekvencie, ktoré sú neskôr preložené na proteín sa nazývajú exóny.
Existuje približne 8 typov intrónov, jeden z nich vyhovuje GT-AG pravidlu. Minimálna dĺžka je 60 bp (musí obsahovať všetky signály potrebné pre orezanie), maximálna dĺžka je neobmedzená. Dĺžka exónov je tiež rôzna: cca 100 – 2000 bp alebo viac. Neexistujú žiadne pravidlá pre distribúciu intrónov [6].

2.6 Alternatívny splicing

Z tých istých transkriptov génov sa tvoria rôzne varianty mRNA. Odhaduje sa, že až 20% ľudských génov má túto vlastnosť, niektoré z nich majú až 64 mRNA variánt. Objav alternatívneho splicingu vyvrátil pôvodnú hypotézu, že z jedného génu sa tvorí vždy jeden proteín. Spôsob spojovania génov môže byť ovplyvnený typom buniek alebo inými okolnosťami.

Napríklad T gén myši obsahuje exón 2 a 3, ktoré sa vzájomne vylučujú, v závislosti na type buniek. Exón 2 je použitý v bunkách hladkého svalstva, zatiaľčo exón 3 je použitý vo všetkých ostatných tkanivách [6].

2.7 Prokaryoty vs. Eukaryoty

Gény majú rozdielne vlastnosti u prokaryotov a eukaryotov. *Prokaryoty* majú vyššiu hustotu génov (85-88% genómu sú kódujúce sekvencie). Jeden promotér zdieľa viac génov súčasne. Prokaryoty majú 1 typ RNA polymeráz a iba niekoľko typov regulátorov. Prokaryoty neobsahujú intróny.

Eukaryoty majú nízku hustotu génov (napr. U človeka je to 5% genómu). Každý gén má svoj vlastný promotér. Eukaryoty majú 3 typy RNA polymeráz a veľké množstvo typov regulátorov. Obsahujú intróny a môže u nich nastať *alternatívny splicing* (Z rovnakých transkriptov génov sa tvoria rôzne varianty mRNA).

2.8 Problémy génovej predikcie

Rastúca dostupnosť obrovských množstiev genetických dát prispela k záujmu o čisto výpočtové metódy pre nájdenie proteíny kódujúcich génov v DNA. Tieto metódy sa musia vysporiadať z rôznymi problémami.

Pri génovej predikcii u *eukaryotov* je potrebné detekovať kde v géne sa nachádzajú *intróny*. To jednak preto, aby vo výsledku neboli zahrnuté do intervalu, na ktorom bol nájdený gén, ale aj preto, aby nebolo predčasne detekované ukončenie génu. Stop kodón v intróne sa totiž neberie do úvahy.

U *eukaryotov* takisto nastáva *alternatívny splicing*. Existujú rôzne modely spájania exónov. Vytvorenie vhodného výpočtového modelu pre spojovanie génov stále zostáva jedným z hlavných problémov algoritmov pre rozpoznávanie génov.

3 Metódy génovej predikcie

Na vyhľadávanie génov existuje mnoho metód a mnoho variánt týchto metód. Prehľad základných metód je spracovaný z prednášky **Rozpoznávaní genů** z predmetu Bioinformatika [6].

3.1 Naivná metóda

Kodón je trojica po sebe idúcich nukleotidov. Existuje 64 možných kodónov. Každá aminokyselina môže byť kódovaná viacerými kodónmi. Každý gén začína *štart kodónom* (AUG) a končí niektorým zo *stop kodónov* (UAA, UAG, UGA). Sekvencia vyskytujúca sa medzi štart a stop kodónom sa nazýva čítací rámec (Open Reading Frame – ORF).

Cieľom metódy je nájdenie všetkých ORF rámcov, čiže nájdenie všetkých sekvencií medzi štart a stop kodónom. Problém je, že nevieme, kde začína kodón a ani či prehľadávame priame alebo reverzné vlákno. Riešením je prehľadanie všetkých 6 možností (3 možnosti začiatku kodónu na priamom aj na reverznom vlákne).

3.2 Naivná metóda s obmedzením dĺžky

Problémom naivnej metódy je, že nevieme, či štart a stop kodóny, ktoré sme našli nie sú v nekódujúcej oblasti. Pokiaľ sa na túto oblasť pozrieme ako na náhodnú sekvenciu kodónov, tak v priemere každých 21 (64/3) sa vyskytuje jeden stop kodón. Gény sú však väčšinou omnoho dlhšie. Riešením je teda hľadať sekvencie medzi štart a stop kodónom dlhšie ako 21 kodónov. Prvé programy na génovú predikciu využívali práve túto metódu. Nie je to však veľmi spoľahlivé, pretože niektoré gény sú kratšie ako 21 kodónov (hlavne gény nervového a imunitného systému) a táto metóda ich nenájde.

3.3 Štatistická metóda

Teória k štatistickej metóde bola prevzatá zo zdroja [2]. Nerovnaké použitie kodónov v kódujúcich oblastiach je univerzálnou vlastnosťou všetkých genómov. Na vytváranie proteínov je použité nerovnomerné množstvo aminokyselín, z toho vyplýva, že je použité nerovnomerné množstvo kodónov pre vytvorenie jednej aminokyseliny. Organizmy toho istého druhu zvyknú mať podobné frekvencie použitia kodónov v kódujúcich oblastiach. Porovnaním frekvencie výskytu kodónov v oblasti s tabuľkou je možné odhadnúť, s akou pravdepodobnosťou sa jedná o kódujúcu oblasť. V praxi je možné túto pravdepodobnosť vypočítať rôznymi spôsobmi. Jedným z nich je pomer logaritmickej pravdepodobnosti.

V tabuľke 2 sú uvedené príslušnosti jednotlivých kodónov k aminokyselinám a percentuálny výskyt jednotlivých kodónov v genóme baktérie E. coli.

	kodón	aminok.	%	kodón	aminok.	%	kodón	aminok.	%	kodón	aminok.	%
	TTT	Phe(F)	1,9	TCT	Ser(S)	1,1	TAT	Tyr(Y)	1,6	TGT	Cys(C)	0,4
	TTC	Phe(F)	1,8	TCC	Ser(S)	1,0	TAC	Tyr(Y)	1,4	TGC	Cys(C)	0,6
	TTA	Leu(L)	1,0	TCA	Ser(S)	0,7	TAA	STOP	0,2	TGA	STOP	0,1
T	TTG	Leu(L)	1,1	TCG	Ser(S)	0,8	TAG	STOP	0,0	TGG	Trp(W)	1,4
	CTT	Leu(L)	1,0	CCT	Pro(P)	0,7	CAT	His(H)	1,2	CGT	Arg(R)	2,4
	CTC	Leu(L)	0,9	CCC	Pro(P)	0,4	CAC	His(H)	1,1	CGC	Arg(R)	2,2
	CTA	Leu(L)	0,3	CCA	Pro(P)	0,8	CAA	Gln(Q)	1,3	CGA	Arg(R)	0,3
C	CTG	Leu(L)	5,2	CCG	Pro(P)	2,4	CAG	Gln(Q)	2,9	CGG	Arg(R)	0,5
	ATT	Ile(I)	2,7	ACT	Thr(T)	1,2	AAT	Asn(N)	1,6	AGT	Ser(S)	0,7
	ATC	Ile(I)	2,7	ACC	Thr(T)	2,4	AAC	Asn(N)	2,6	AGC	Ser(S)	1,5
	ATA	Ile(I)	0,4	ACA	Thr(T)	0,1	AAA	Lys(K)	3,8	AGA	Arg(R)	0,2
A	ATG	Met(M)	2,6	ACG	Thr(T)	1,3	AAG	Lys(K)	1,2	AGG	Arg(R)	0,2
	GTT	Val(V)	2,0	GCT	Ala(A)	1,8	GAT	Asp(D)	3,3	GGT	Gly(G)	2,8
	GTC	Val(V)	1,4	GCC	Ala(A)	2,3	GAC	Asp(D)	2,3	GGC	Gly(G)	3,0
	GTA	Val(V)	1,2	GCA	Ala(A)	2,1	GAA	Glu(E)	4,4	GGA	Gly(G)	0,7
G	GTG	Val(V)	2,4	GCG	Ala(A)	3,2	GAG	Glu(E)	1,9	GGG	Gly(G)	0,9
	T			C			A			G		

Tabuľka 2: Početnosť kodónov v genóme baktérie E. coli

3.3.1 Početnosť kodónov

Početnosť kodónov vyjadruje aké je percento jednotlivých kodónov v danej sekvencii. Tieto hodnoty sa porovnávajú s početnosťou kodónov vo vybranom genóme.

Nech $F(c)$ je frekvencia (pravdepodobnosť) kodónu c v génoch skúmaného druhu (inými slovami: F je tabuľka početnosti kodónov). Ak máme postupnosť kodónov $C=C_1C_2\ldots C_m$ a predpokladáme nezávislosť susedných kodónov, tak

$$P(C)=F(C_1)F(C_2)\cdots F(C_m)$$

je pravdepodobnosť nájdenia sekvencie kodónov C, pričom C kóduje proteín.

Príklad:

ak S je sekvencia S = `TTATTT` a rátame od rámca 1, vznikne sekvencia $C_1^1=TTA, C_2^1=TTT$

Potom

$$P^1(S)=P(C^1)=F(TTA)F(TTT)$$

Dosadením potrebných hodnôt z tabuľky 2 dostaneme

$$P^1(S)=P(C^1)=0.01*0.019=0.00019$$

Nech $F_0(c)$ je frekvencia kodónu c v nekódujúcej oblasti:

$$P_0(S)=P_0(C)=F_0(C_1)F_0(C_2)\cdots F_0(C_m)$$

je pravdepodobnosť nájdenia sekvencie S ak C je nekódujúce. Za predpokladu, že náhodný model kódujúcej DNA, $F_0(c)=1/64=0.0156$ pre všetky kodóny a P_0 pre vyššie uvedenú sekvenciu kodónov C, tak:

$$P_0(C)=0.0156*0.0156=0.000244$$

Logaritmická pravdepodobnosť, že S je v rámci 1 kódovacia oblasť LP^1, je

$$LP^1(S)=\log(0.00019/0.000244)=\log(0.778688)=-0.108636$$

Logaritmická pravdepodobnosť pomerov pre S kódujúce v rámcoch 2 a 3 (LP^2 a LP^3) sa vypočíta obdobným spôsobom. V uvedenom príklade je logaritmická pravdepodobnosť vyššia ako nula, čo vyjadruje kódovaciu oblasť, zatiaľčo nižšia ako nula vyjadruje nekódovaciu oblasť.

V praxi sa táto metóda využíva skôr na objavenie (väčšinou malých) kódovacích oblastí vo veľkých génových sekvenciách. Typickým postupom je vypočítať štatistickú pravdepodobnosť v posuvných oknách a zaznamenať hodnotu pre každé z týchto okien. Takto vznikne priebeh, v ktorom maximá môžu poukazovať na kódovacie oblasti a minimá nekódovacie oblasti.

3.3.2 Použitie aminokyselín

Ide o štatistiku toho, koľko je v sekvencii percent kodónov kódujúcich konkrétne aminokyseliny.

$$F_A(c)=\sum_{c^0\equiv c} F(c')$$

c' znamená kodón synonymický ku kodónu c

$$P_A^i(S)=P_A(C^i)=F_A(C_1^i)F_A(C_2^i)\cdots F_A(C_m^i)$$

P_a^i je pravdepodobnosť nájdenia sekvencie aminokyselín, ktorá vznikne preložením sekvencie S v rámci i, ak S je kódovacie v rámci i. V nekódujúcej DNA budeme predpokladať pravdepodobnosť , že každá aminokyselina je proporčná počtu synonymických kodónov kódujúcich aminokyselinu. $F_{A0}(c)=n_c/64$ kde n_c , je počet kodónov synonymických k c . Môžeme vypočítať $P_{A0}(S)$. A potom z $P_A^i(S)$ a $P_{A0}(S)$ vieme vypočítať logaritmický pomer použitia aminokyseliny.

Príklad:

ak S je sekvencia S =`TTATTT` a rátame od rámca 1, vznikne sekvencia $C_1^1=TTA, C_2^1=TTT$. Kodón `TTA` kóduje *Leucín*, ten môžu kódovať kodóny `TTA,TTG,CTT,CTC,CTA,CTG` (6 synonymických kodónov, čiže $n_c=6$). Z toho vyplýva:

$$F_A(TTA)=F_A(TTA)+F_A(TTG)+F_A(CTT)+F_A(CTC)+F_A(CTA)+F_A(CTG)$$
$$F_A(TTA)=0.01+0.011+0.01+0.009+0.003+0.052=0.095$$
$$F_{A0}(TTA)=\frac{n_c}{64}=\frac{6}{64}=0.09375$$

Kodón TTT kóduje *Fenylalanín*, ten môžu kódovať kodóny TTT a TTC (2 synonymické kodóny, čiže $n_c=2$). Z toho vyplýva:

$$F_A(TTT)=F_A(TTT)+F_A(TTC)$$
$$F_A(TTT)=0.019+.0.018=0.037$$
$$F_{A0}(TTT)=\frac{n_c}{64}=\frac{2}{64}=0.03125$$

$$P_A^1(TTATTT)=P_A(TTATTT^1)=F_A(TTA^1)F_A(TTT^1)$$
$$P_A^1(TTATTT)=0.095*0.037=0.003515$$
$$LP^1(TTATTT)=\log(\frac{0.095*0.037}{0.09375*0.03125})=\log(1.199786)=0.079104$$

3.3.3 Preferencia kodónov

Je možné spraviť aj štatistiku na odmeranie nerovnomerného použitia synonymických kodónov (kodóny, ktoré kódujú tú istú aminokyselinu). Z tabuľky početnosti kodónov vieme vypočítať relatívnu pravdepodobnosť každého synonymického kodónu pre danú aminokyselinu. Nech $F_R(c)$ je relatívna pravdepodobnosť, že v kódujúcej oblasti bude spomedzi kodónov synonymických k c použitý páve kodón c.

$$F_R(C)=F\frac{(C)}{\sum_{c^0\equiv c}F(c')}$$

$$P_R^i(S)=P_R(C^i)=F_R(C_1^i)F_R(C_2^i)\cdots F_R(C_m^i)$$

P_R^i je pravdepodobnosť, že sekvencia S kóduje dané aminokyseliny v rámci i . Budeme predpokladať, že v nekódujúcej DNA nie je žiadna preferencia medzi synonimickými kodónmi pre danú aminokyselinu. Pravdepodobnosť kodónu c v nekódujúcej oblasti DNA vypočítame podielom čísla 1 a počtu kodónov, ktoré môžu kódovať danú aminokyselinu. $F_{R0}=1/n_c$. Z $P_R^i(S)$ a $P_{R0}(S)$ vypočítame logaritmický pomer preferencie kodónov.

Príklad:
ak S je sekvencia S =TTATTT a rátame od rámca 1, vznikne sekvencia $C_1^1=TTA, C_2^1=TTT$. Kodón TTA kóduje *Leucín*, ten môžu kódovať kodóny TTA,TTG,CTT,CTC,CTA,CTG (6 synonymických kodónov, čiže $n_c=6$). Z toho vyplýva:

$$F_R(TTA)=\frac{F(TTA)}{F(TTA)+F(TTG)+F(CTT)+F(CTC)+F(CTA)+F(CTG)}$$

$$F_R(TTA)=\frac{0.01}{0.01+0.011+0.01+0.009+0.003+0.052}=0,105263$$

$$F_{R0}(TTA)=\frac{1}{6}=0,166666$$

Kodón `TTT` kóduje *Fenylalanín*, ten môžu kódovať kodóny `TTT` a `TTC` (2 synonimické kodóny, čiže $n_c=2$). Z toho vyplýva:

$$F_R(TTT)=\frac{F(TTT)}{F(TTT)+F(TTC)}$$

$$F_R(TTT)=\frac{0,019}{0,019+0,018}=0,513513$$

$$F_{R0}(TTT)=\frac{1}{2}=0,5$$

Pravdepodobnosť,že práve sekvencia S kóduje dané aminokyseliny vypočítame:

$$P_R^i(S)=P_R(C^i)=F_R(C_1^i)F_R(C_2^i)\cdots F_R(C_m^i)$$

$$P_R^i(TTATTT)=P_R(TTATTT^1)=F_R(TTA^1)F_R(TTT^1)=0.105263+0.513513=0.618776$$

Logaritmická pravdepodobnosť, že S je v rámci 1 kódovacia oblasť LP^1 , je

$$LP^1(TTATTT)=\log\left(\frac{0.105263*0.513513}{0.166666*0.5}\right)=\log(0,6486496)=-0,187989$$

3.3.4 Súvislosti

I keď použitie aminokyselín a preferencia kodónov nesú dosť informácií o kódovaní, ani jedno z týchto kritérií nie je tak závažné ako početnosť kodónov. V skutočnosti je početnosť kodónov kompozíciou **použitia aminokyselín** a **preferencie kodónov**. Táto skutočnosť je zrejmá aj zo vzorcov:

$$F(c)=F_A(c)F_R(c)$$

$$F_0(c)=F_{A0}(c)F_{R0}(c)$$

dostaneme

$$P^i(S)=P_A^i(S)P_R^i(S)$$

$$P_0(S)=P_{A0}(C)P_{R0}(C)$$

pre sekvenciu S v rámci i , z čoho vyplýva

$$LP^i(S)=LP_A^i(S)+LP_R^i(S)$$

To značí, že početnosť kodónov je sumou použitia aminokyselín a preferencie kodónov.

3.3.5 Použitie Hexamerov

hexamer	výskyt(%)
AAAAAA	0.107809
AAAAAC	0.076021
AAAAAG	0.186379
AAAAAT	0.001050
AAAACA	0.022042
AAAACC	0.030139
AAAACG	0.085618
AAAACT	0.020692
...	...

Tabuľka 3: Časť tabuľky hexamerov u baktérie E. coli

Aj odchýlky v rozložení oligonukleotidov iných než kodóny môžu byť použité na rozlíšenie kódovacích a nekódovacích oblastí. Odchýlky v použití hexamerov (dvojíc kodónov) sú najzávažnejším z doteraz spomínaných štatistických faktorov. Vypočítame ich rovnako ako početnosť kodónov. Tabuľka početnosti hexamerov $F(h_i)(i=1,\cdots,4096)$ pre práve skúmaný druh organizmu je vypočítaná apriori. Pravdepodobnosť sekvencie hexanukleotidov, $H=H_{1,}H2,\ldots,H_m$ v kódovacej oblasti sekvencie je $P(H)=F(H_1)F(H_2)\cdots F(H_m)$.

Ak P_0 je distribučná funkcia pravdepodobnosti, logaritmický pomer LP je možné vypočítať ako vo vyššie uvedených prípadoch. Testovacia sekvencia môže byť rozobratá na šesť rôznych sekvencií hexamerov miesto troch, a preto sa počítajú pravdepodobnosti pre 6 rámcov ($LP^i, i=1\cdots 6$).

Príklad:
ak H je sekvencia H = AAAAAAAAAAAG a rátame od rámca 1, vznikne sekvencia $H_1^1=AAAAAA, H_2^1=AAAAAG$
Potom

$$P^1(H)=P(AAAAAAAAAAAG^1)=F(AAAAAA)F(AAAAAG)$$

Dosadením potrebných hodnôt z tabuľky 3 dostaneme

$$P^1(H)=P(C^1)=0.00107809*0.00186379=0.0000020093333611$$

Frekvencia postupnosti hexamerov v nekódovacej oblasti:

$$P_0(H)=\frac{1}{4096}*\frac{1}{4096}=0.000000059604644775390625$$

Logaritmická pravdepodobnosť, že H je v rámci 1 kódovacia oblasť LP^1 , je

$$LP^1(S)=\log(P^1(H)/P_0(H))=\log(33.711)=1.527771$$

3.4 Pozične špecifické matice

Pozične špecifická matica (PWM – Position Weight Matrix) je bežne používanou reprezentáciou vzorov v biologických sekvenciách.
PWM je matica hodnôt skóre, ktoré dáva váhu akémukoľvek danému podreťazcu pevnej dĺžky. Má jeden riadok pre každý symbol danej abecedy (v našom prípade A,C,T,G) a jeden stĺpec pre každú pozíciu vzoru.

Skóre je pridelené maticou podreťazcu $s=(s_j)_{j=1}^{N}$ je definované ako súčet $\sum_{j=1}^{N} m_{s_j,\,j}$, kde j vyjadruje polohu v podreťazci, s_j je symbol na pozícii j v podreťazci, a $m_{\alpha,\,j}$ je skóre v riadku α , stĺpci j matice. Inými slovami, PWM skóre je sumou pozične špecifických skóre pre každý symbol v podreťazci.

3.4.1 Základná PWM s logaritmickými pravdepodobnosťami

PWM predpokladá nezávislosť medzi pozíciami vo vzore, keďže počíta skóre pre každú pozíciu nezávisle od symbolov na ostatných pozíciách. Skóre podreťazca zarovnané s PWM môže byť interpretované ako logaritmická pravdepodobnosť podreťazca pod multinomickým rozložením pravdepodobnosti. Keďže každý stĺpec definuje logaritmické pravdepodobnosti pre každý symbol, kde suma pravdepodobností v stĺpci je rovná číslu 1, PWM korešponduje s multinomickým rozložením pravdepodobnosti. Jednotlivé skóre v PWM môžu byť tiež chápané vo fyzickej štruktúre ako suma väzbových energií pre všetky nukleotidy zarovnané s PWM [11].

3.4.2 Príklad použitia PWM: Predikcia začiatku translácie

Teória k tejto metóde bola popísaná na základe zdroja [9]. Niektoré nukleotidy sa okolo štart kodónu (ATG) vyskytujú častejšie, ak v tejto oblasti začína translácia. V tejto metóde je potrebné vziať niekoľko známych génov z toho istého druhu organizmu a spraviť štatistiku kodónov v okolí začiatku translácie.

Spravil som štatistiku, ktoré kodóny sa v baktérii E. coli zvyknú vyskytovať v okolí štart kodónu (ATG), ak tam začína gén. V tabuľke 4 sú v percentách uvedené pravdepodobnosti výskytu daného kodónu na konkrétnej pozícii voči začiatku štart kodónu.

pozícia	-4	-3	-2	-1	0,1,2	3	4
A	37,08	40,86	26,21	25,92	ATG	50,12	35,47
C	21,86	18,14	27,54	27,26	ATG	16,52	28,16
T	24,11	16,95	32,84	30,6	ATG	14,89	19,33
G	16,95	24,06	13,41	16,23	ATG	18,47	17,04

Tabuľka 4: Výskyt báz v okolí štart kodónu pri začatí translácie u baktérie E. Coli

Ako táto metóda funguje je ukázané na nasledujúcom príklade. Máme dve sekvencie obsahujúce štart kodón (ATG). U ktorej z nich je väčšia pravdepodobnosť začiatku translácie?
CACCATGGC

TCGAATGTT

Matematický model: $F_i(x)$ je frekvencia $X(A, C, T, G)$ na pozícii i.

Je treba vypočítať skóre daného reťazca pomocou vzorca $\sum_i \log(F_i(X)/0.25)$

CACCATGGC	TCGAATGTT
log(21,86/25) + log(40,86/25) + log(27,54/25) +	log(24,11/25) + log(18,14/25) + log(13,41/25) +
log(27,26/25) + log(18,74/25) + log(28,16/25) =	log(25,92/25) + log(14,89/25) + log(19,33/25) =
-0.0582 + 0.2133 + 0.0420 + 0.0375 -0.1251	-0.0157 - 0.1393 - 0.2705 + 0.0156 - 0.2250
+0.0516	-0.1117
= 0.1611	= -0.7466

Ak je skóre vyššie ako 0, predpokladáme, že na tomto mieste začína translácia. Čím vyššie je skóre, tým vyššia je pravdepodobnosť, že je predikcia správna. V tomto prípade, môžeme predpokladať, že prvá sekvencia (CACCATGGC) naozaj začína transláciu a druhá (TCGAATGTT) nie.

3.5 Skryté Markove Modely (Hidden Markov Models – HMM)

Skrytý Markov model je stavový generatívny model, ktorý prevádza symboly konečnou abecedou. Môže sa nachádzať v jednom z množiny diskrétnych stavov. Prechody medzi stavmi sú dané podmienenou pravdepodobnosťou. Pre hodnoty pravdepodobnosti platí, že sú väčšie ako 0 a že ich súčet je rovný číslu 1. Väčšinou hodnoty pravdepodobnosti modelu (hodnoty prechodov a emisné pravdepodobnosti) nepoznáme a musíme ich získať trénovaním modelu na známej množine vzorov výstupných sekvencií a sekvencií stavov – trénovacia množina. Existuje celá rada metód pre trénovanie HMM. Najčastejšie sa používa Baum-Welchov algoritmus.

Aplikácia HMM na rozpoznávanie génov pozostáva z troch krokov. (1) Najprv vykonáme ručnú identifikáciu jednotlivých stavov a prechodov na základe znalosti o štruktúre génu. (2) Potom trénujeme model. Emisné a prechodové pravdepodobnosti určíme na základe sekvencií so známymi génmi. (3) Na záver použijeme vytrénovaný model na analýzu neznámej sekvencie a nájdeme najpravdepodobnejšie rozdelenie rozdelenie na exóny, intróny a medzigénovú oblasť pomocou *Viterbiho algoritmu* [5].

4 Presnosť génovej predikcie

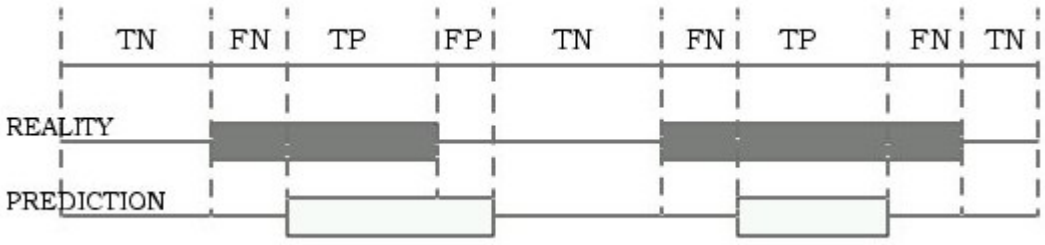
Hodnotenie presnosti algoritmu na detekciu génov je veľmi dôležité pre vývojárov aj pre používateľov týchto algoritmov, pretože práve na základe výsledkov programov na génovú predikciu sú vykonávané experimenty, na ktoré je treba vynaložiť veľa úsilia a zdrojov. Pre vývojárov je dôležité, aby vedeli v akom stave je ich program a kde ho môžu vylepšiť. Hodnotenie presnosti bude mať význam aj v rámci tejto bakalárskej práce, pretože na základe tohto hodnotenia budem určovať, ktoré metódy sú lepšie a aké parametre u danej metódy dávajú najlepší výsledok. Hodnotenie metód som spracoval na základe [1].

4.1 Meranie presnosti predikcie

Presnosť jednotlivých algoritmov je možné hodnotiť na rôznych úrovniach:

1. úroveň báz
2. úroveň exónov

4.1.1 Úroveň báz



Určuje presnosť algoritmu podľa kódovacíh a nekódovacích báz. U každej bázy sa porovnáva, či patrí do niektorého génu a či ju metóda označila ako patriacu do génu.

True positives (TP) sú bázy, ktoré patria do nejakého génu a metóda ich označila ako patriace do génu. *True negatives (TN)* sú bázy, ktoré nepatria do žiadneho génu a metóda ich označila ako nepatriace do žiadneho génu. *False positives (FP)* sú bázy, ktoré nepatria do žiadneho génu, ale boli metódou označené ako patriace do nejakého génu. *False negatives (FN)* sú bázy, ktoré patria do nejakého génu, ale metóda ich označila, ako bázy nepatriace do žiadneho génu.

Citlivosť je schopnosť metódy nájsť všetky bázy skutočného génu. *Špecifickosť* je schopnosť metódy neoznačiť bázy, ktoré do génu nepatria za súčasť génu.

Sn (Sensitivity): Citlivosť: $Sn = \frac{TP}{TP + FN}$

Sp (Specificity): Špecifickosť: $Sp = \frac{TN}{TN + FP}$

Príklad:
DNA sekvencia je 300 báz dlhá, prvá báza má číslo 0. Gén sa nachádza na bázach 189-255. Metóda detekovala gén na bázach 180-248 (Reálne organizmy majú samozrejme väčší počet génov a dlhšie DNA sekvencie, ale pre ilustráciu postačí aj takýto príklad). Aká je citlivosť a špecifickosť metódy?

Interval báz	**0-179**	**180-188**	**189-248**	**249-255**	**256-299**
Počet báz	180	9	60	7	44
Realita					
Predpoveď					
TP/TN/FP/FN	TN	FP	TP	FN	TN

Tabuľka 5: Ukážkový príklad hodnotenia metódy na úrovni báz

$TP = 60$

$TN = 180 + 44 = 224$

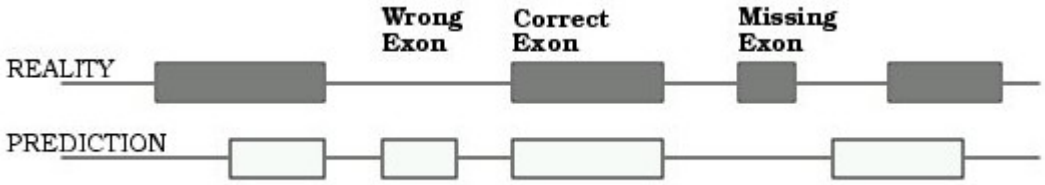
$FP = 9$

$FN = 7$

$Sn = \frac{60}{60 + 7} = 0.895 \qquad Sp = \frac{224}{224 + 9} = 0.961$

Daná metóda má na úrovni báz približne 89% citlivosť a 96% špecifickosť.

4.1.2 Úroveň exónov



Určuje presnosť algoritmu s ohľadom na presnú predikciu začiatku a konca exónu. *Správne detekované exóny (Correct Exon)* sú exóny, ktoré boli nájdené metódou a prekrývajú skutočnú polohu exónu z rovnakej alebo väčšej časti ako daný prah α. *Chýbajúce exóny (Missing Exon)* sú reálne exóny, ktoré metóda ani sčasti nedetekovala. *Nesprávne detekované exóny (Wrong Exon)* sú exóny, ktoré boli detekované na mieste, kde sa v skutočnosti žiadny exón nenachádza

Citlivosť určuje aké percento reálnych exónov metóda skutočne našla. *Špecifickosť* určuje aké percento z detekovaných exónov sú správne detekované exóny.

Citlivosť: $Sn = \frac{počet\ Správne\ detekovaných\ exónov}{počet\ Skutočných\ exónov}$

Špecifickosť: $Sp = \frac{počet\ Správne\ detekovaných\ exónov}{počet\ detekovaných\ exónov}$

Príklad:
V DNA vybraného organizmu sa nachádza 4200 exónov. Metóda detekovala 6500 exónov, z toho 3500 detekovala správne. Aká je citlivosť a špecifickosť metódy?

$$Sn = \frac{počet\ Správne\ detekovaných\ exónov}{počet\ Skutočných\ exónov} = \frac{3500}{4200} = 0.8333$$
$$Sp = \frac{počet\ Správne\ detekovaných\ exónov}{počet\ detekovaných\ exónov} = \frac{3500}{6500} = 0.5385$$

Metóda má približne 83% citlivosť a 53% špecifickosť.

4.2 Implementácia hodnotenia metód

Metódy na predikciu som programoval tak, aby na štandardný výstup vypísali nájdené gény vo formáte:

```
<znamienko vlákna (priame: +, reverzné: -)> <začiatok génu (číslo bázy)> <koniec génu (číslo bázy)>
```

Príklad výpisu:

```
+   29   98
-   107   500
-   169   187
+   189   255
-   229   364
+   336   2799
…
```

Tento výstup porovnávam so skutočným zoznamom génov v rovnakom formáte a zisťujem zhody. Metódu potom hodnotím *na úrovni exónov*. Prokaryotické organizmy (teda aj E. coli) nemajú intróny, takže hodnotenie na úrovni exónov je v ich prípade hodnotenie na úrovni génov. Toto hodnotenie vyjadruje koľko génov trafila metóda presne, čo je zaujímavý údaj, avšak hodnotenie týmto spôsobom je príliš hrubé, pretože ak metóda detekuje začiatok alebo koniec génu inde než v skutočnosti je (hoci len o niekoľko báz), gén sa počíta za nenájdený.
Preto uvádzam aj hodnotenie na úrovni báz, kde je vidieť, na koľko je metóda presná z hľadiska celkovej plochy. U oboch hodnotení uvádzam citlivosť (senzitivitu) a špecifickosť metódy. Niekedy medzi sebou porovnávam aj rôzne variácie tej istej metódy

5 Experimenty

5.1 Modelový organizmus: Baktéria E. coli

E. coli je často používaná ako modelový organizmus v štúdiách mikrobiológie. Bakteriálna konjugácia (proces prenosu genetického materiálu medzi bakteriálnymi bunkami) bola prvý krát objavená použitím E. coli ako modelovej baktérie. E. coli sa na skúmanie konjugácie dodnes používa ako primárny model. E. coli bola zásadnou súčasťou prvých experimentov na pochopenie genetiky baktériofága. Seymour Benzer použil E. coli a baktériofága T4 na pochopenie topografie génovej štruktúry. Pred týmto výskumom sa nevedelo či má gén lineárnu alebo rozvetvenú štruktúru. E. coli bola jedným z prvých organizmov, ktorým zmapovali genóm. Kompletný genóm E. coli K12 bol uverejnený v roku 1997 [8].

Ako modelový organizmus na testovanie metód génovej predikcie používam E. coli K12. Genóm E. coli K12 je dlhý 4 639 657 bázových párov, z toho približne 50% tvoria gény.

5.2 Naivná metóda

Naivnú metódu v najzákladnejšej forme som implementoval ako detektor ORF rámcov bez akýchkoľvek obmedzení. ORF (čítací rámec) je sekvencia nukleotidov medzi štart a stop kodónom. Nie v každom rámci sa nachádza gén a tak je možné predpokladať, že naivná metóda označí za gény aj sekvencie, kde gény nie sú.

Naivná metóda		
Úroveň	Exóny	Bázy
Citlivosť	73,31%	98,91%
Špecifickosť	3,85%	33,49%

Tabuľka 6: Naivná metóda

Naivná metóda detekovala presne približne 73% génov baktérie E. Coli a na úrovni báz 98,91% obsahu. Metóda má výbornú citlivosť. Problém je veľmi nízka špecifickosť. 96% detekovaných génov bolo false positives (alebo boli nájdené, ale nie presne). Na úrovni báz bolo nesprávne označených vyše 66% obsahu. Pre lepšie výsledky bude potrebné obmedziť počet false positives.

5.3 Naivná metóda s obmedzením dĺžky

Ak sa na nekódujúcu oblasť pozrieme ako na náhodnú sekvenciu kodónov, potom v priemere každých (64/3) = 21 kodónov vyskytuje jeden stop kodón. Gény sú častokrát oveľa dlhšie. Upravil

som naivnú metódu, tak aby hľadala len oblasti medzi štart a stop kodónom, kotré sú dlhšie ako 21 kodónov. Výskyt stop kodónu, skôr ako 21 kodónov po štart kodóne bude považovaný za náhodu. Vyskúšal som aj variáciu metódy s vyšším obmedzením dĺžky ako na 21 kodónov.

Naivná metóda s obmedzením dĺžky nad 21 kodónov mala o niečo menej false positives a teda lepšiu špecifickosť (4,01%), ale aj slabšiu citlivosť (61%).

Pomocou skriptu som vypočítal priemernú dĺžku jednotlivých génov v baktérii E. Coli a navrhol rôzne obmedzenia dĺžky pre naivnú metódu. Najkratší gén v baktérii E. Coli má 15 kodónov, najdlhší má 7279 kodónov, priemerná dĺžka génu je približne 320 kodónov. Otestoval som metódu s obmedzením dĺžky pre 21, 30, 50, 100, 200 a 300 kodónov.

Ako je možné vidieť v tabuľke 7, na úrovní exónov jemne vzrastá citlivosť, čo je pravdepodobne spôsobené tým, že v niektorých prípadoch by naivná metóda ukočíla gén na skoršom stop kodóne, než mala - táto situácia pri obmedzení dĺžky nenastala. Preto obmedzenie dĺžky jemne znížilo počet false positives. S jemným zvyšovaním špecifickosti však rapídne klesá citlivosť a od obmedzenia dĺžky na 100 kodónov klesá aj špecifickosť.

Na úrovni báz sa citlivosť drží na tej istej úrovni a so zvyšovaním dĺžky klesá špecifickosť. To je dané tým, že s obmedzením dĺžky metóda nájde bázy, ktoré do génu patria, ale odignoruje stop kodóny, ktoré sú bližšie k štart kodónu než zvolené obmedzenie, preto má z hľadiska báz vysoký počet false positives a teda nízku špecifickosť.

Naivná metóda

bez obmedzení		
Úroveň	Exóny	Bázy
Citlivosť	73,31%	98,91%
Špecifickosť	3,85%	33,49%
obmedzenie dĺžky na 21 kodónov		
Úroveň	Exóny	Bázy
Citlivosť	61,95%	99,41%
Špecifickosť	4,01%	20,35%
obmedzenie dĺžky na 30 kodónov		
Úroveň	Exóny	Bázy
Citlivosť	58,16%	99,53%
Špecifickosť	4,07%	16,62%
obmedzenie dĺžky na 50 kodónov		
Úroveň	Exóny	Bázy
Citlivosť	51,31%	99,68%
Špecifickosť	4,19%	11,30%
obmedzenie dĺžky na 100 kodónov		
Úroveň	Exóny	Bázy
Citlivosť	38,27%	99,88%
Špecifickosť	4,15%	5,34%
obmedzenie dĺžky na 200 kodónov		
Úroveň	Exóny	Bázy
Citlivosť	22,86%	99,95%
Špecifickosť	3,62%	1,75%
obmedzenie dĺžky na 300 kodónov		
Úroveň	Exóny	Bázy
Citlivosť	11,38%	99,97%
Špecifickosť	2,35%	0,95%

Tabuľka 7: Naivná metóda s obmedzením dĺžky

5.4 Predikcia začiatku translácie

Metóda predikcie translácie		
Úroveň	Exóny	Bázy
Citlivosť	52,20%	87,82%
Špecifickosť	6,37%	67,86%

Tabuľka 8: Metóda predikcie translácie

Z hľadiska exónov je táto metóda len o trocha efektívnejšia ako naivná metóda s obmedzením dĺžky, ktorá pri podobnej citlivosti dosahovala o 2 percentá väčšiu špecifickosť. Pri hodnotení z hľadiska báz je však rozdieľ oveľa viditeľnejši. Za cenu mierneho zníženia citlivosti (približne o 10%), táto metóda dosiahla viac než dvojnásobnú úroveň špecifickosti (z 33% na 67%). To znamená že táto metóda má približne o tretinu menej false positives než naivná metóda.

5.5 Štatistická metóda

5.5.1 Početnosť kodónov

Štatistická metóda

Okno = 60, prah = 0		
Úroveň	Exóny	Bázy
Citlivosť	73,41%	98,00%
Špecifickosť	4,90%	37,45%
Okno = 60, prah = 1		
Úroveň	Exóny	Bázy
Citlivosť	73,22%	98,51%
Špecifickosť	6,01%	41,87%
Okno = 60, prah = 2		
Úroveň	Exóny	Bázy
Citlivosť	71,66%	97,81%
Špecifickosť	8,06%	49,68%
Okno = 60, prah = 3		
Úroveň	Exóny	Bázy
Citlivosť	67,45%	96,22%
Špecifickosť	11,69%	60,69%
Okno = 60, prah = 4		
Úroveň	Exóny	Bázy
Citlivosť	56,36%	92,00%
Špecifickosť	16,64%	73,00%
Okno = 60, prah = 5		
Úroveň	Exóny	Bázy
Citlivosť	40.3%	84,38%
Špecifickosť	21,13%	85,73%
Okno = 60, prah = 6		
Úroveň	Exóny	Bázy
Citlivosť	20,93%	70,01%
Špecifickosť	18,93%	94,75%
Okno = 60, prah = 7		
Úroveň	Exóny	Bázy
Citlivosť	6,68%	46,30%
Špecifickosť	11,02%	99,05%

Tabuľka 9: Štatistická metóda - početnosť kodónov

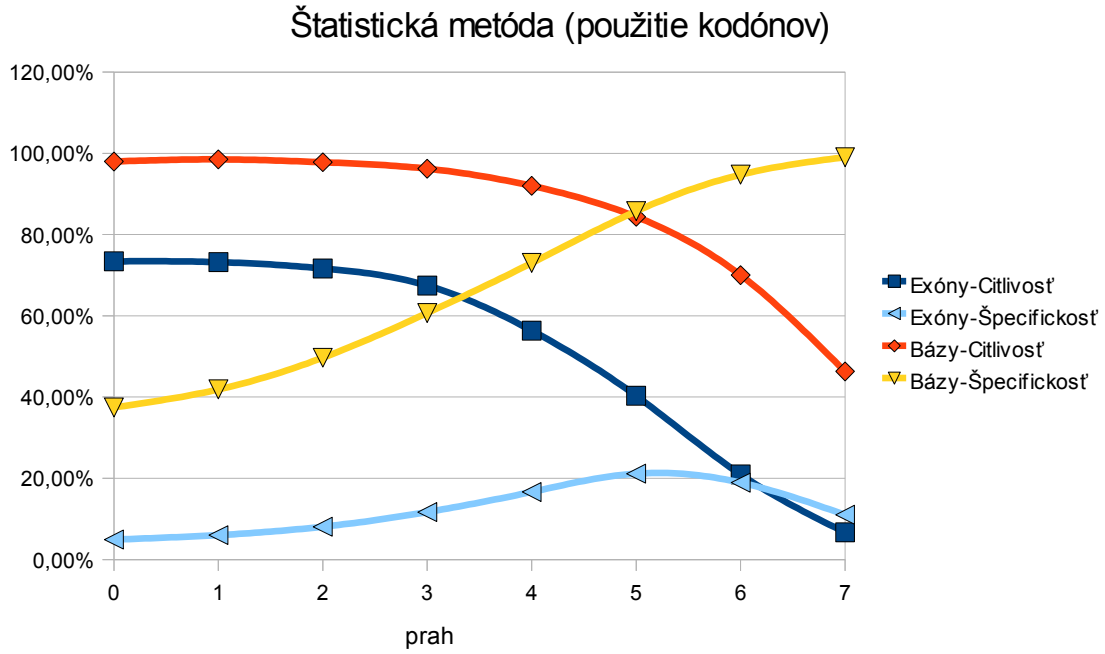
Uvedená teória hovorí o tom, ako štatisticky ohodnotiť postupnosť znakov (báz) a získať tak skóre, ktoré vyjadruje pravdepodobnosť, že daná postupnosť je v kódujúcej oblasti. Avšak nie je jasné ako štatistickú metódu implementovať. Aby som ukázal účinnosť štatistickej metódy, vyhnem sa zatiaľ kombinovaniu s inými metódami. Napriek tomu zostáva mnoho možností a parametrov implementácie tejto metódy.

Jedným z typických použití štatistickej metódy je robenie štatistiky v okne danej veľkosti, ktoré sa posúva o daný počet báz. Tento spôsob iste má využitie pri predikcii DNA u eukaryotov, kde výkyvy v štatistike jednotlivých oblastí môžu naznačovať striedanie intrónov a exónov.

Prokaryotické organizmi však nemajú intróny a stop kodón teda nemôže byť v nekódujúcej oblasti. U prokaryotov v drivivej väčšine prípadov stop kodón zastaví transláciu a jediným problémom zostáva určiť, či nájdený štart kodón transláciu začína. Je to podobné ako predikcia translácie pomocou báz v okolí štart kodónu s tým rozdielom, že v tomto prípade bude sledovaná sekvencia vybranej dĺžky za štart kodónom, kde by sa mohol nachádzať potenciálny gén. Ak štatistika ukáže, že sekvencia je kódujúcou oblasťou, gén bude označený až po najbližší stop kodón. Ak štatistika ukáže, že ide o nekódujúcu oblasť, bude štart kodón ignorovaný a program bude pokračovať v hľadaní ďalších štart kodónov.

Aj v tomto prípade však zostávajú dva premenlivé parametre. Prvým je dĺžka sekvencie za štart kodónom, ktorá bude skúmaná. Druhým je prah, teda hodnota skóre, od ktorej už bude oblasť považovaná za kódujúcu. Skóre sekvencie nie je tak jasné číslo ako napríklad percentuálna hodnota a závisí od dĺžky sekvencie, z ktorej sa robí štatistika. Preto pre rôzne veľkosti skúmanej sekvencie budú rôzne prahy, pri ktorých bude metóda najefektívnejšia.

Najprv som vyskúšal skúmať sekvenciu 60 báz za štart kodónom a sledoval som výsledky pri rôznych hodnotách prahu.



Z grafu vidíme, že čím vyšší dáme prah, tým lepšia je citlivosť a horšia špecifickosť. Je to logické, pretože, ak sprísnime podmienky metódy, nájdeme menej false positives, ale aj skutočných génov nájdeme menej. Z hľadiska exónov je efektívne zvyšovať prah maximálne po hodnotu 6. Pri prahu 6 klesá nielen citlivosť ale aj špecifickosť.

Z hľadiska báz je to lepšie. Pri prahu o hodnote 5 je veľmi podobná citlivosť a špecifickosť, oboje majú hodnotu okolo 85%. Pri prahu o hodnote 7 je špecifickosť 99,05%, ale citlivosť je len 46,3%, V tomto prípade metóda neoznačí takmer žiadne false positives, ale objaví menej ako polovicu génov (myslené z hľadiska plochy, ktoré gény pokrývajú), čo je opačná situácia ako pri prahu 0.

5.5.2 Početnosť hexamerov

Urobil som štatistiku počtu hexamerov v génoch baktérie E. Coli. Namerané údaje som uložil a použil v programe, ktorý vyhľadáva gény na základe štatistiky hexamerov. Tabuľku pre jej obrovský rozsah (4096 objektov) neuvádzam. Je možné ju nájsť v zdrojovom kóde, ktorý je súčasťou prílohy.
Program funguje na rovnakom princípe ako pri sledovaní početnosti kodónov s tým rozdielom, že sleduje početnosť hexamerov a posúva sa o tri políčka. Je to preto, že každý kodón (okrem krajných) môže tvoriť dvojicu s kodónom pred ním aj s kodónom za ním.

Použitie hexamerov			
	Okno = 120, prah = 0		
	Úroveň	Exóny	Bázy
	Citlivosť	71,66%	96,41%
	Špecifickosť	23,18%	72,63%
	Okno = 120, prah = 1		
	Úroveň	Exóny	Bázy
	Citlivosť	69,89%	95,50%
	Špecifickosť	26,99%	76,57%
	Okno = 120, prah = 2		
	Úroveň	Exóny	Bázy
	Citlivosť	67,63%	94,39%
	Špecifickosť	31,00%	80,42%
	Okno = 120, prah = 3		
	Úroveň	Exóny	Bázy
	Citlivosť	64,56%	93,04%
	Špecifickosť	35,05%	84,18%
	Okno = 120, prah = 4		
	Úroveň	Exóny	Bázy
	Citlivosť	60,51%	87,58%
	Špecifickosť	38,26%	85,37%
	Okno = 120, prah = 5		
	Úroveň	Exóny	Bázy
	Citlivosť	55,57%	88,83%
	Špecifickosť	40,47%	90,55%

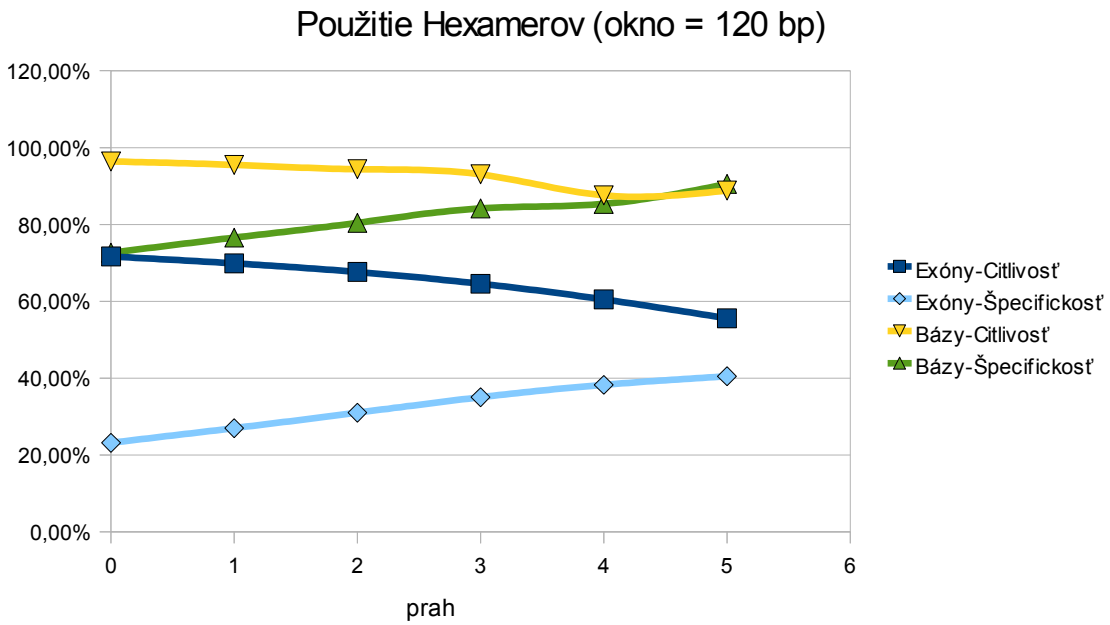
Tabuľka 10: Použirie hexamerom s oknom 120bp

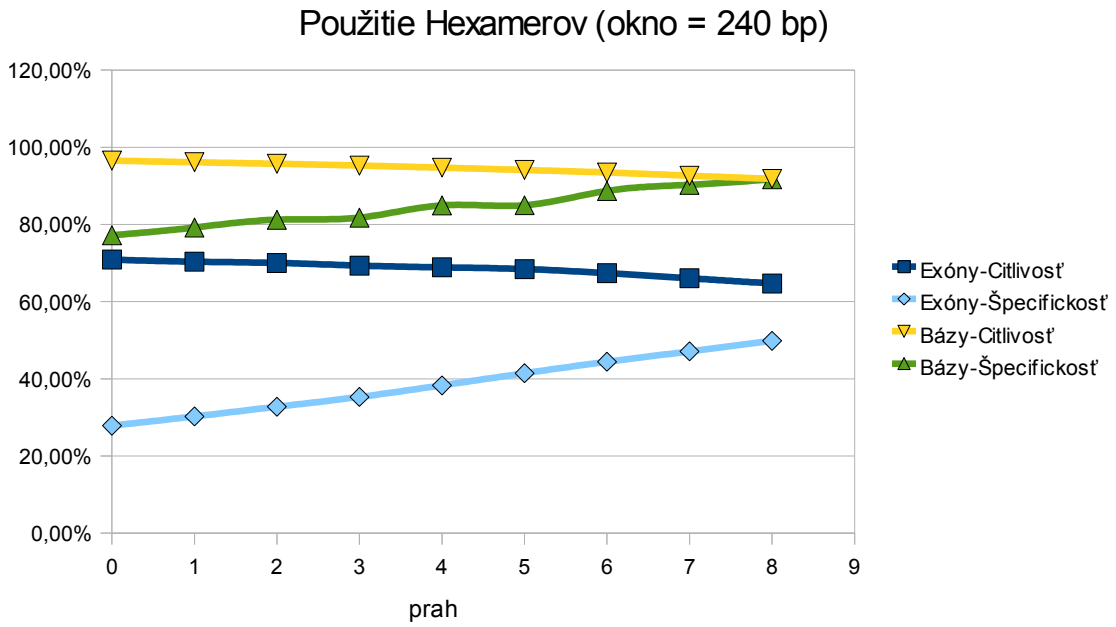
Použitie hexamerov			
	Okno = 240, prah = 0		
	Úroveň	Exóny	Bázy
	Citlivosť	70,89%	96,56%
	Špecifickosť	27,86%	77,19%
	Okno = 240, prah = 1		
	Úroveň	Exóny	Bázy
	Citlivosť	70,36%	96,13%
	Špecifickosť	30,25%	79,21%
	Okno = 240, prah = 2		
	Úroveň	Exóny	Bázy
	Citlivosť	70,03%	95,73%
	Špecifickosť	32,77%	81,25%
	Okno = 240, prah = 3		
	Úroveň	Exóny	Bázy
	Citlivosť	69,35%	95,28%
	Špecifickosť	35,30%	81,78%
	Okno = 240, prah = 4		
	Úroveň	Exóny	Bázy
	Citlivosť	68,89%	94,73%
	Špecifickosť	38,31%	84,96%
	Okno = 240, prah = 5		
	Úroveň	Exóny	Bázy
	Citlivosť	68,45%	94,12%
	Špecifickosť	41,44%	84,99%
	Okno = 240, prah = 6		
	Úroveň	Exóny	Bázy
	Citlivosť	67,38%	93,48%
	Špecifickosť	44,41%	88,71%
	Okno = 240, prah = 7		
	Úroveň	Exóny	Bázy
	Citlivosť	66,07%	92,64%
	Špecifickosť	47,10%	90,29%
	Okno = 240, prah = 8		
	Úroveň	Exóny	Bázy
	Citlivosť	64,70%	91,83%
	Špecifickosť	49,83%	91,74%

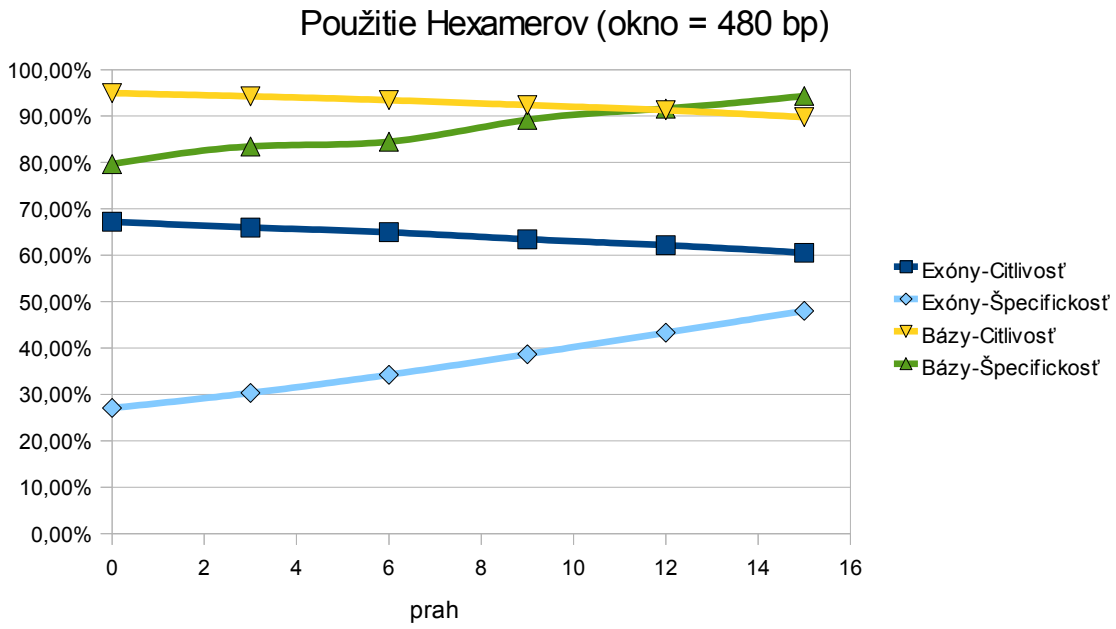
Tabuľka 11: Použitie hexamerov s oknom 240bp

Použitie hexamerov			
	Okno = 480, prah = 0		
	Úroveň	Exóny	Bázy
	Citlivosť	67,24%	94,95%
	Špecifickosť	27,06%	79,63%
	Okno = 480, prah = 3		
	Úroveň	Exóny	Bázy
	Citlivosť	65,98%	94,26%
	Špecifickosť	30,31%	83,47%
	Okno = 480, prah = 6		
	Úroveň	Exóny	Bázy
	Citlivosť	64,96%	93,43%
	Špecifickosť	34,23%	84,47%
	Okno = 480, prah = 9		
	Úroveň	Exóny	Bázy
	Citlivosť	63,45%	92,37%
	Špecifickosť	38,65%	89,20%
	Okno = 480, prah = 12		
	Úroveň	Exóny	Bázy
	Citlivosť	62,16%	91,26%
	Špecifickosť	43,31%	91,62%
	Okno = 480, prah = 15		
	Úroveň	Exóny	Bázy
	Citlivosť	60,53%	89,76%
	Špecifickosť	48,00%	94,34%

Tabuľka 12: Použitie hexamerov s oknom 480bp







Štatistickú metódu s použitím hexamerov som testoval pre veľkosti okna 120, 240 a 480 bázových párov. Varianta so šírkou okna 120bp dosahovala pri svojom optimálnom prahu citlivosť a špecifickosť približne 90%. Varianta so šírkou okna 240bp dosahovala na úrovní báz podobnú, avšak o niečo vyššiu špecifickosť. Na úrovni exónov dosahovala varianta 240bp pri tomto istom prahu približne o 3% lepšiu špecifickosť a takmer o 10% lepšiu citlivosť. Variantu 240bp je teda možné považovať za efektívnejšiu ako variantu 120bp. Varianta 480bp dosahovala na úrovni báz podobnú efektivitu ako metóda 240bp, na úrovni exónov však mala o niečo horšie výsledky. Z toho je možné usúdiť, že optimálna veľkosť okna pre túto metódu sa nachádza v intervale 240-480bp. Zisťovanie najoptimálnejšej varianty tejto metódy by bolo časovo príliš náročné a odkláňalo by sa od témy tejto práce. Z vyskúšaných variant je najlepšia varianta 240bp pri prahu 8.

5.6 Signály

5.6.1 Gilbertov a Pribnow box

Na základe percent u jednotlivých kodónov v konsenzus sekvencii (varianta Lisser-Margalit) som vytvoril pozične špecifickú maticu. Percentá pre danú bázu na konkrétnej pozícii odlišnú než vzorovú som vypočítal rovnomerným rozdelením zostávajúcich percent medzi zostávajúce bázy.

pozícia	1	2	3	4	5	6
A	10,33	7	13	56	15,33	54
C	10,33	7	13	14,66	54	15,33
T	69	79	13	14,66	15,33	15,33
G	10,33	7	61	14,66	15,33	15,33

Tabuľka 13: matica pravdepodobností pre Pribnow box (TTGACA)

dĺžka medzery	pravdepodonbnosť
16	17
17	43
18	17

Tabuľka 14: veľkosť medzery medzi PB a GB

pozícia	1	2	3	4	5	6
A	7,66	76	13,33	61	56	6
C	7,66	8	13,33	13	14,66	6
T	77	8	60	13	14,66	82
G	7,66	8	13,33	13	14,66	6

Tabuľka 15:matica pravdepodobností pre Gilbertov box (TATAAT)

V mojej implementácii sledujem jednotlivé šestice báz pre výskyt Pribnowho boxu (TATAAT) a porovnávaním s maticou. Každej pozícii prideľujem skóre, ktoré určuje, či je pravdepodobné, že na danom mieste je Pribnow box. K pozíciam, u ktorých je skóre vyššie ako nula hľadám Gilbertov box vo vzdialenostiach 16,17 a 18 báz. Všetky tri možnosti predstavujú potenciálny promotér (Gilbertov box a k nemu prislúchajúci Pribnow box). Z možností označím tú, ktorá získala najvyššie celkové skóre. Ak je toto skóre vyššie ako nula predpokladám, že na danom mieste sa nachádza promotér.

5.6.2 Zakomponovanie metódy do programu

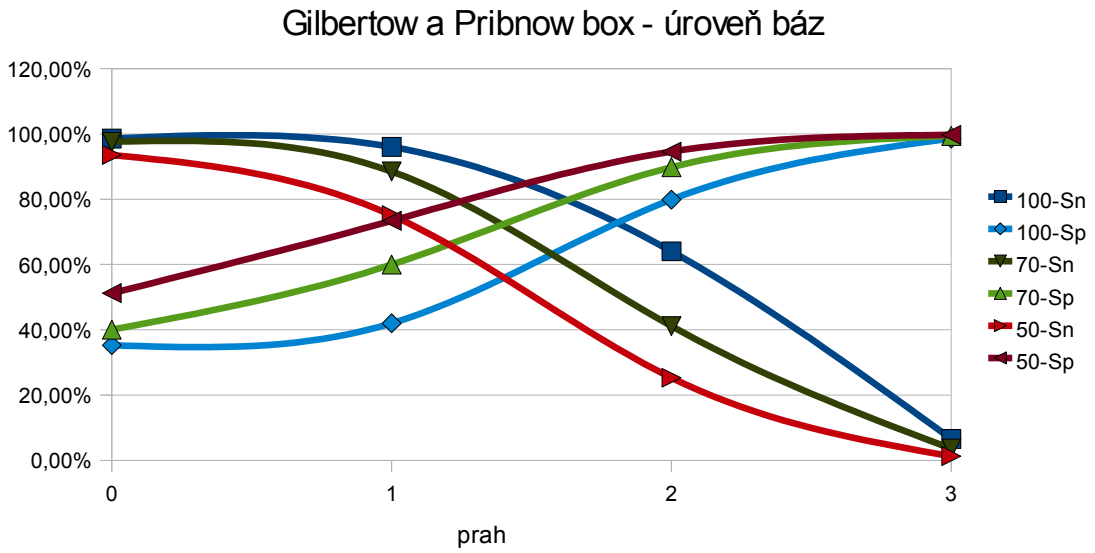
Promotéry	Vzdialenosť	100		70		50		30	
		Prah = 0		Prah = 0		Prah = 0		Prah = 0	
	Úroveň	Exóny	Bázy	Exóny	Bázy	Exóny	Bázy	Exóny	Bázy
	Citlivosť	71,47%	98,59%	66,49%	97,60%	54,76%	93,55%	7,21%	36,53%
	Špecifickosť	3,87%	35,29%	3,94%	40,05%	4,11%	51,22%	3,41%	92,08%
		Prah = 1		Prah = 1		Prah = 1		Prah = 1	
	Úroveň	Exóny	Bázy	Exóny	Bázy	Exóny	Bázy	Exóny	Bázy
	Citlivosť	65,21%	96,00%	46,72%	88,51%	32,89%	75,09%	2,51%	16,09%
	Špecifickosť	4,03%	41,97%	4,29%	59,97%	4,55%	73,39%	3,12%	97,00%
		Prah = 2		Prah = 2		Prah = 2		Prah = 2	
	Úroveň	Exóny	Bázy	Exóny	Bázy	Exóny	Bázy	Exóny	Bázy
	Citlivosť	27,75%	64,03%	15,08%	41,15%	8,66%	25,30%	0,32%	2,71%
	Špecifickosť	4,88%	79,89%	5,15%	89,79%	5,48%	94,58%	3,26%	99,68%
		Prah = 3		Prah = 3		Prah = 3		Prah = 3	
	Úroveň	Exóny	Bázy	Exóny	Bázy	Exóny	Bázy	Exóny	Bázy
	Citlivosť	2,32%	6,53%	1,46%	3,57%	0,65%	1,26%	0,02%	0,06%
	Špecifickosť	5,31%	98,61%	7,38%	99,40%	6,96%	99,72%	12,50%	99,99%

Tabuľka 16: Gilbertow a Pribnow box pri rôznych hodnotách prahu a vzdialenosti od štart kodónu

Postup je podobný ako při štatistickej metóde. Keď je nájdený štart kodón, preskúma sa oblasť před štart kodónom a zistí sa, s akou pravdepodobnosťou sa tam nachádza dvojica Gilbertovho a Pribnowho boxu. Máme dva parametre, ktoré je možné upravovať. Vzdialenosť před štart kodónom, ktorá bude braná do úvahy a prah skóre, teda od akého skóre vyššie bude predpokladané, že na danom mieste sa dvojica boxov naozaj nachádza a že štart kodón je začiatkom génu. Metódu som testoval pre vzdialenosti 100, 70, 50, 30 a prahy 0, 1, 2, 3.

Z pohľadu exónov má najlepšiu špecifickosť metóda, ktorá sleduje 30 báz před štart kodónom. Tá má však veľmi nízku citlivosť a je preto nepoužiteľná. Ostatné metódy majú medzi sebou veľmi podobnú špecifickosť a aj citlivosť, pričom najlepšiu citlivosť má metóda, ktorá skúma 100 báz před štart kodónom. Hodnotenie podľa exónov ale nie je dostatočne smerodajné, pretože pracuje len s presnými detekciami daného génu.

Hodnotenie metódy na úrovni báz poskytuje smerodajnejšie výsledky. Aj tu vidíme, že varianta s dĺžkou 30 báz ma veľmi slabú citlivosť a nie je veľmi použiteľná. Zvyšné metódy majú lepšie výsledky a všetky 3 vyzerajú použiteľne. Pre rozhodnutie, ktorá varianta je najlepšia bude dobré znázorniť špecifickosť aj citlivosť variánt s dĺžkami 50, 70 a 100 v jednom grafe.



Z grafu je vidno, že so vzrastajúcim prahom klesá citlivosť a vzrastá špecifickosť. Je otázne, ktorá z týchto vlastností je dôležitejšia. Chceme, aby metóda našla, čo najviac génov za cenu väčšieho množstva false positives, alebo budeme požadovať, aby metóda bola čo najpresnejšia, ale zmierime sa a tým, že niektoré gény nebudú detekované? Osobne si myslím, že obe vlastnosti sú rovnako dôležité a metóda je teda najlepšia tam, kde má podobnú citlivosť a špecifickosť. V grafe som zvolil podobné odtiene farieb pre danú dĺžku, aby bolo vidieť, při ktorej hodnote má daná varianta metódy rovnakú citlivosť a špecifickosť. Z grafu je možné odhadnúť optimálny prah pre každú variantu.

varianta	prah	Citlivosť = špecifickosť
50	≈1,0	≈75%
70	≈1,4	≈75%
100	≈1,8	≈75%

Tabuľka 17: Tabuľka optimálnych prahov pre rôzne vzdialenosti

Ako je vidno z tabuľky 17, všetky tri metódy sú takmer rovnako efektívne, len každá při inom prahu. Ak berieme do úvahy aj rýchlosť vykonania algoritmu, tak je samozrejme najefektívnejšia varianta, ktorá pozerá 50 znakov pred štart kodónom, pretože to trvá kratší čas.

5.7 Väzobné miesta ribozómov

Táto metóda vyhľadáva väzobné miesta ribozómov (ribosomial binding sitest).Pri implementácii programu som použil Konsenzus sekvenciu uvedenú v zdroji [3]. V tabuľke 18 sú uvedené pravdepodobnosti získané z tejto sekvencie.

pozícia	1	2	3	4	5	6
A	38,00%	55,00%	9,00%	10,00%	60,00%	27,00%
C	31,00%	15,00%	7,00%	7,00%	10,00%	7,00%
T	7,00%	17,00%	11,00%	5,00%	14,00%	16,00%
G	24,00%	14,00%	73,00%	75,00%	13,00%	46,00%

Tabuľka 18: tabuľka pravdepodobností pre väzobné miesta ribozómov

Ak program nájde štart kodón skúma všetky možné šestice v oblasti 3-20 bázových párov pred štart kodónom. Ak aspoň jedna zo šestíc má skóre pre konsenzus sekvenciu nad určeným prahom, nájdený štart kodón je označený ako začiatok génu.

Skúšal som rôzne prahy, až kým som sa nedostal k hodnote prahu, kde by citlivosť a špecifickosť na úrovni báz mali približne rovnakú hodnotu. Výsledky je možné vidieť v tabuľke 19.

Ribosomial binding sites			
	prah = 0		
	Úroveň	Exóny	Bázy
	Citlivosť	61.62%	94,30%
	Špecifickosť	5.08%	55,06%
	prah = 1		
	Úroveň	Exóny	Bázy
	Citlivosť	31,41%	56,00%
	Špecifickosť	11,43%	90,93%
	prah = 0,5		
	Úroveň	Exóny	Bázy
	Citlivosť	47,56%	83,29%
	Špecifickosť	6,78%	74,39%
	prah = 0,75		
	Úroveň	Exóny	Bázy
	Citlivosť	38,49%	71,06%
	Špecifickosť	8,51%	84,15%
	prah = 0,6		
	Úroveň	Exóny	Bázy
	Citlivosť	43,33%	78,40%
	Špecifickosť	7,41%	78,91%

Tabuľka 19: Detekcia väzobných miest ribozómov pre rôzne prahy

5.8 Porovnanie metód

	Exóny			Bázy	
	Sn	Sp	(Sp+Sn)/2	Sn	Sp
Naivná metóda	0,73	0,04	0,39	0,99	0,33
Metóda predikcie translácie	0,52	0,06	0,29	0,88	0,68
Väzobné miesta ribozómov (RBS)	0,43	0,07	0,25	0,78	0,79
Pribnow a Gilbertow box	0,33	0,05	0,19	0,75	0,73
Štatistická metóda (početnosť kodónov)	0,40	0,21	0,31	0,84	0,86
Štatistická metóda (hexamery)	0,65	0,50	0,57	0,92	0,92

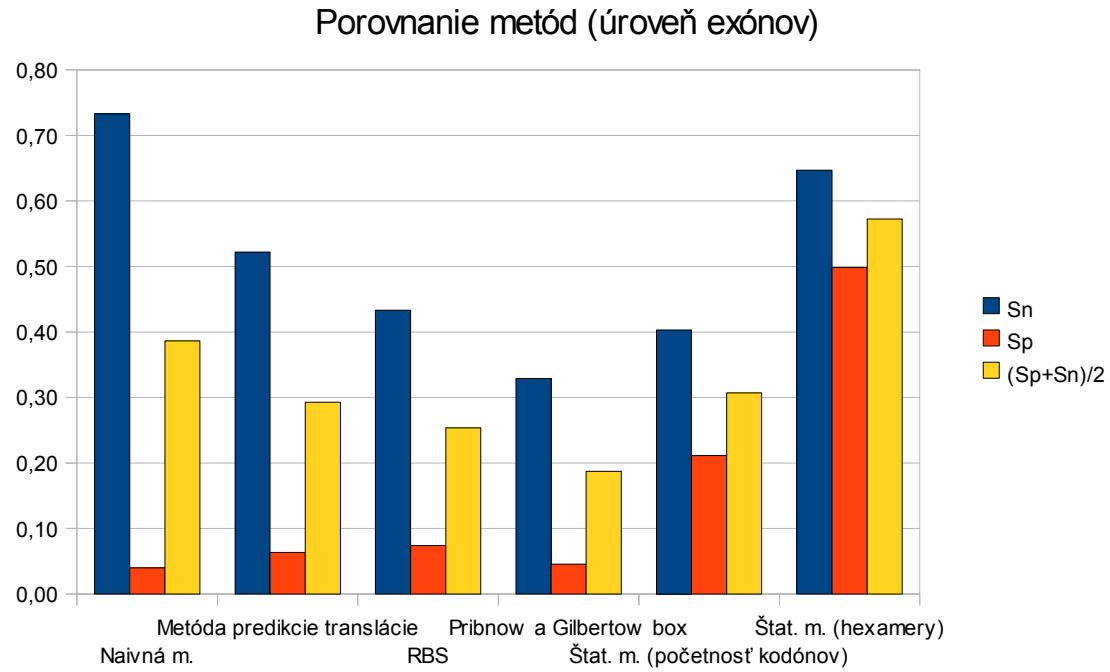
Tabuľka 20: Porovnanie metód

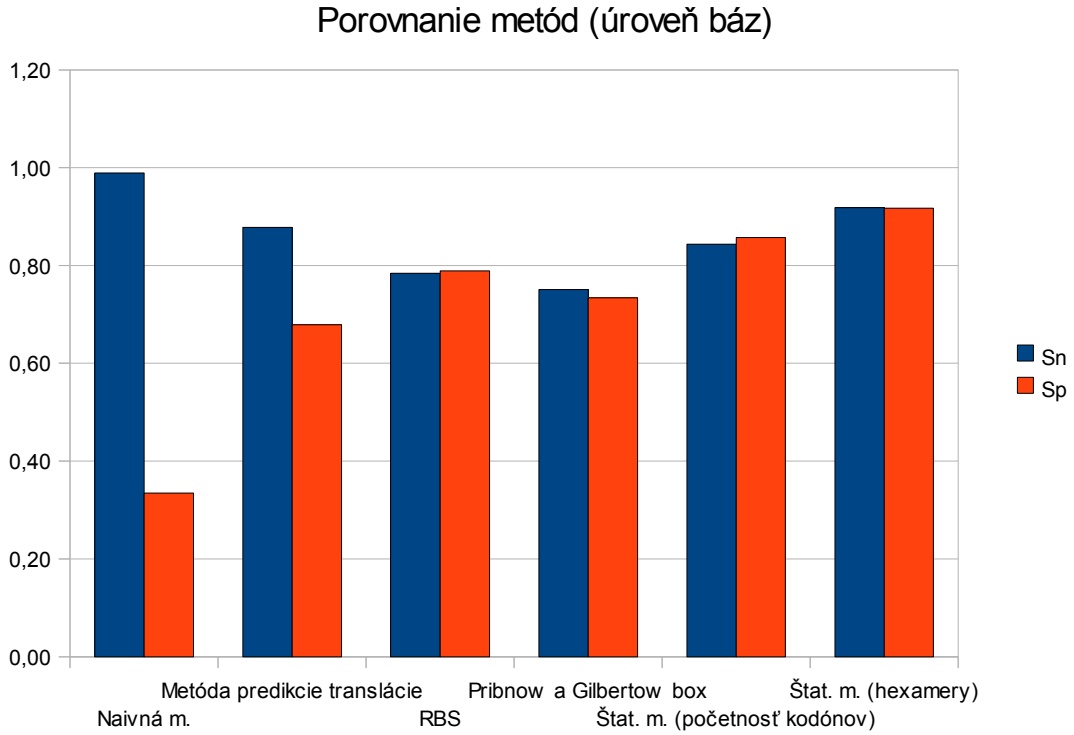
Na záver porovnám všetky metódy. Každá metóda má mnoho variánt, preto som sa snažil vždy vybrať najoptimálnejšiu variantu metódy pre porovnanie. Metódy porovnávam na úrovni exónov aj na úrovni báz.

Na úrovni exónov je najlepšia štatistická metóda s použitím hexamerov, nasleduje štatistická metóda s počítaním početnosti kodónov. Ostatné metódy majú na úrovni exónov slabú citlivosť.

Na úrovni báz je opäť najlepšia štatistická metóda s použitím hexamerov s citlivosťou aj špecifickosťou nad 90%. Štatistická metóda s meraním početnosti kodónov dosahuje tiež veľmi dobré výsledky s citlivosťou aj špecifickosťou približne 85%. Metódy využívajúce signály majú podobnú efektivitu, hoci metóda RBS je o čosi lepšia ako metóda pre rozpoznávanie promotérov (Pribnow a Gilbertow box). Podobne účinná je aj metóda predikcie translácie. Je to zaujímavé pretože táto metóda napevno sleduje iba niekoľko báz pred a za štart kodónom a metódy využívajúce signály hľadajú konsenzus sekvenciu v celom vymedzenom rozsahu. Najslabšia je samozrejme naivná metóda kvôli nízkej citlivosti.

V prípade implementácie metódy na vyhľadávanie génov, kde by sa tieto jednotlivé metódy kombinovali by bolo najvhodnejšie použiť štatistiku hexamerov v kombinácii s metódami pre detekciu signálov.


Porovnanie metód (úroveň exónov)
0,80
0,70
0,60
0,50
0,40
0,30
0,20
0,10
0,00
Sn
Sp
(Sp+Sn)/2
Naivná m.
Metóda predikcie translácie
RBS
Pribnow a Gilbertow box
Štat. m. (početnosť kodónov)
Štat. m. (hexamery)


Porovnanie metód (úroveň báz)
1,20
1,00
0,80
0,60
0,40
0,20
0,00
Sn
Sp
Naivná m.
Metóda predikcie translácie
RBS
Pribnow a Gilbertow box
Štat. m. (početnosť kodónov)
Štat. m. (hexamery)

6 Záver

V tejto práci som sa zaoberal témou detekcie génov v DNA sekvenciách prokaryotických organizmov. Popísal som dôležité pojmy týkajúce sa tejto problematiky a prehľad metód. Ako modelový organizmus pre testovanie som zvolil baktériu E. coli K12. Vybrané metódy som implementoval, otestoval pre rôzne parametre a zhodnotil ich efektivitu. V praxi sa tieto metódy metódy kombinujú, ja som ich však testoval samostatne, aby bolo možné ich porovnať. Keby som mal v práci pokračovať, vidím dva rôzne smery, ktorými by sa mohla práca uberať. (1) Pokračoval by som vo forme prehľadu a testovania metód samostatne s rozšírením metód na eukaryotické organizmy, kde by bolo treba riešiť nové problémy: oddelenie exónov od intrónov, alternatívny splicing, iné signály ako u prokaryotov a podobne. (2) Druhou možnosťou by bolo zostať pri zameraní na prokaryotické organizmy a začať metódy kombinovať a pracovať tak na jednej finálnej metóde. V tomto prípade by bolo treba starostlivo zvoliť ako na seba metódy naviazať, aké im dať prahy a akú váhu dať každej metóde.

Literatúra

[1] Guigó, R.: *Accuracy of Gene prediction methods* [online]. [cit. 2011-5-9]. Dostupné z WWW: <http://genome.crg.es/courses/genefinding/T6/index.html>

[2] Guigó. R.: *DNA Composition, Codon Usage and Exon Prediction* [online]. Barcelona, 2000. [cit. 2011-5-9]. Dostupné z WWW: <http://www.pdg.cnb.uam.es/cursos/FVi2001/GenomAna/GeneIdentification/SearchContent/main.html>

[3] Hayes, W. S. - Borodovsky, M.: *Deriving Ribosomial Binding Site (RBS) statistical models from unannotated DNA sequences and the use of the RBS model for N-terminal prediction* [online]. [cit. 2011-5-9]. Dostupné z WWW: <http://helix-web.stanford.edu/psb98/hayes.pdf>

[4] Kodíček, M.: *Biochemické pojmy - výkladový slovník*. VŠCHT v Praze, 2007, ISBN 978-80-7080-669-2.

[5] Majoros, W. H. - Ohler, U.: *Advancing the State of the Art in Computational Gene Prediction*. Center for Bioinformatics and Computational Biology, 2007.

[6] Martínek, T.: *Rozpoznávání genů*. Vysoké Učení Technické v Brně - Fakulta informačních technologií.

[7] Oliviera et. al.: *Ribosome binding site recognition using neural networks*. Genetics and Molecular Biology, vol. 27, no.4., São Paulo, 2007, ISSN 1415-4757.

[8] *Escherichia coli* [online]. Wikipedia. [cit. 2011-5-9] Dostupné z WWW: <http://en.wikipedia.org/wiki/Escherichia_coli >

[9] Zhai, C.: *Gene Prediction: Statistical Methods* [online]. University of Illinois, 2005. [cit. 2011-5-9]. Dostupné z WWW: <http://sifaka.cs.uiuc.edu/course/498cxz05f/ppt/genstat.ppt>

[10] *Genetics Home Reference* [online]. [cit. 2011-5-9]. Dostupné z WWW: <http://ghr.nlm.nih.gov/handbook>,

[11] *Position-specific Scoring Matrix* [online]. Wikipedia. [cit. 2011-5-9]. Dostupné z WWW: <http://en.wikipedia.org/wiki/Position-specific_scoring_matrix>

[12] *The Bacterial Promoter* [online]. [cit. 2011-5-9]. Dostupné z WWW: <http://www.mun.ca/biochem/courses/3107/Topics/promoter_bacterial.html>

Zoznam príloh

Príloha 1. CD/DVD so zdrojovými textami
Príloha 2. Tabuľka súborov (skripty, pomocné dáta...)

Tabuľka súborov

Názov súboru	Popis funkcie
analyzaGenomu.py	Skript, ktorý zanalyzuje početnosť kodónov vo vybranom genóme
analyzaGenomuHexamer.py	Skript, ktorý zanalyzuje početnosť hexamerov vo vybranom genóme
analyzaGenomuStart.py	Skript, ktorý zanalyzuje použitie báz v okolí štart kodónov na začiatku génov
dlzkaGenov.py	Skript, ktorý zanalyzuje minimálnu,maximálnu a priemernú dĺžku génov
naivna.py	Implementácia naivnej metódy
naivnaPredikciaStartu.py	Implementácia metody pre predikciu začiatku translácie
porovnanie.py	Skript, ktorý ohodnotí metódu na úrovni exónov
porovnanieBaz.py	Skript, ktorý ohodnotí metódu na úrovni báz
PribnowBoxORF.py	Implementácia metódy pre detekcie Pribnowho a Gilbertovho boxu
RBS.py	Implementácia metody pre predikciu ribosomial binding sites
statORF.py	Implementácia štatistickej metody pre početnosť kodónov
statORFhexa.py	Implementácia štatistickej metody pre početnosť hexamerov
dnaecoli.txt	Textový súbor, v ktorom je zapísaná DNA baktérie E. coli K12 (bez hlavičky)
ecoliGeny.txt	Textový súbor, v ktorom sú vypísané gény baktérie E. coli K12 v uvedenom formáte. Používa sa na porovnávanie výstupu s implementovanými metódami

Príklad použitia 1:

```
naivna.py -dna=dnaecoli.txt > vystupNaivnej.txt
porovnanie.py -real=ecoliGeny.txt -method=vystupNaivnej.txt
porovnanieBaz.py -dna=dnaecoli.txt -real=ecoliGeny.txt
-method=vystupNaivnej.txt
```

Príklad použitia 2:

```
RBS.py -dna=dnaecoli.txt > vystupRBS.txt
porovnanie.py -real=ecoliGeny.txt -method=vystupRBS.txt
porovnanieBaz.py -dna=dnaecoli.txt -real=ecoliGeny.txt
-method=vystupRBS.txt
```

Poznámka:

Návod na použitie konkrétneho skriptu je možné získať použitím parametru -help (napr. `porovnanie.py -help`)